UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Diversidade e Evolução de Elementos de Transposição em *Drosophila*

Adriana Ludwig

Tese de Doutorado submetida ao Programa de Pós-graduação em Genética e Biologia Molecular como requisito parcial para obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular)

Orientador: Élgion Lúcio da Silva Loreto

Co-orientadora: Vera Lúcia da Silva Valente Gaiesky

Porto Alegre, Abril de 2010.

Este trabalho foi realizado no Laboratório de *Drosophila* do Departamento de Genética da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil, com bolsa e recursos do CNPq. Parte do trabalho também foi realizada no laboratório do Dr. Harmit S. Malik, na Divisão de Ciência Básica do Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA, Estados Unidos, com bolsa de doutorado sanduíche pelo CNPq e recursos do NSF (US National Science Foundation).

In the future, attention undoubtedly will be centered on the genome, with greater appreciation of its significance as a highly sensitive organ of the cell that monitors genomic activities and corrects common errors, senses unusual and unexpected events, and responds to them, often by restructuring the genome. We know about the components of genomes that could be made available for such restructuring. We know nothing, however, about how the cell senses danger and instigates responses to it that often are truly remarkable.

(Barbara McClintock, 1984)

Dedico essa tese ao Newton e aos meus pais, Euclésio e Terezinha

Agradecimentos

Ao meu orientador, Professor Elgion, pela confiança depositada em mim, pela orientação, compreensão, conselhos, apoio e amizade. Por compartilhar sua linha de pesquisa e despertar meu interesse pelos elementos de transposição. Por compartilhar conhecimentos e valores, que foram fundamentais para meu desenvolvimento profissional e pessoal.

À minha co-orientadora, Professora Vera, pelo seu carinho, incentivo, e ajuda. Pela sua preocupação em proporcionar aos alunos boas condições e um ótimo ambiente de trabalho.

Ao Dr. Harmit, meu supervisor do doutorado sanduíche, por ter me dado a oportunidade de trabalhar em seu laboratório, em um centro de pesquisa de excelência nos Estados Unidos, e ter me propiciado um grande crescimento profissional. Pela paciência, compreensão e incentivo que me deram forças para continuar em momentos difíceis.

Ao Elmo e à Ellen, por toda a ajuda sempre. Por tornarem mais fácil a vida dos alunos do PPGBM.

Aos meus colegas colaboradores, Maríndia, Newton e Nina, pela oportunidade prazerosa e produtiva de ter trabalhado com vocês. À Maríndia, muito obrigada por ter ajudado a Dirleane.

À Dirleane, minha querida orientada! Tem sido um grande aprendizado e satisfação trabalhar com você. Sinto por tê-la "abandonado" durante meu doutorado sanduíche, mas sei que você ficou em boas mãos!

Aos colegas do Lab de *Drosophila* da UFRGS, Dirleane, Maríndia, Nina, Juliana WG, Juliana C, Marícia, Mário, Hermes, Cleverton, Gilberto, Gisele, Carol, pelos momentos alegres e descontraídos que passamos no laboratório e fora dele.

Aos colegas do Malik Lab, Aida, Ben, Celia, Danielle, Emily B, Emily H, Erik, Josh, Kevin, Marianne, Matt, Maulik, Mia, Nels, Nitin, Ray, Scott, pela paciência e ajuda e pelo exemplo de ambiente acadêmico. Em especial ao Ray que se tornou um grande amigo.

Aos profissionais do FHCRC, Carolyn Goard (Lab Administrator), Michele Karantsavelos (Manager of Graduate Education), Stephanie Otis (Human Resource Assistant) e à Meire Marcilene (SEBIE-CNPq), pela grande ajuda e pela disponibilidade em esclarecer todas minhas dúvidas durante o período de doutorado sanduíche.

À Nanna e Hula, Kim e Ani, Olivia (OP!) e Rex, pela grande amizade que construímos, pelos ótimos momentos de descontração, por terem sido minha família em Seattle. À OP e a Nanna em especial por terem me ajudado muito quando cheguei em Seattle e sempre. Tive muita sorte de ter convivido com vocês!

À Kat, Savannah, Ceci, Jackie, Amber, Andres e Ray, pela amizade e carinho, e por ótimos momentos de descontração e alegria.

Às minhas queridas e eternas amigas, Maríndia, Liz e Nina, pela nossa grande amizade que nasceu no LabDros e que continua sempre. Não existem palavras para descrever meus sentimentos. Obrigada pela ajuda, conselhos, puxões de orelha, pelos inúmeros momentos de alegria que passamos juntas, enfim, obrigada pela amizade. À Nina e a Dharma, pelas constantes visitas que tornavam aquele apartamento muito mais alegre.

Ao Daniel, Cris, Felipe e Lu. Que saudades! Foi muito bom ter morado e convivido com vocês!

À Paula, por ser uma ótima companheira de apartamento, pela sua amizade e carinho e por agüentar todas minhas bagunças por tanto tempo.

Aos meus pais e minhas irmãs Neiva e Aline, pelo amor, carinho, ajuda, pela compreensão de minha ausência. Aos meus pais, pelos valores de vida que me passaram e que me acompanham sempre. Obrigada pelo grande incentivo que sempre me deram. À minha vó Eli, pelo amor, carinho e pela ajuda que sempre me deu.

À minha segunda família, Daca, Luiz, D. Geny e S. Crescêncio (*in memorian*), pelo grande carinho, amizade e ajuda. À Daca, pelo seu exemplo de dedicação e amor a família.

Ao Newton. Meu maior admirador e incentivador. Nesses 10 anos já enfrentamos muitos obstáculos para seguir nossos sonhos. A distância insistiu em ser constante. Seu grande amor, carinho, ajuda e apoio foram muito importantes para eu chegar até aqui. Obrigada por todo o incentivo que me deste, mesmo que isso significasse ficarmos bem longe. Ter o teu apoio foi muito importante para eu seguir meus sonhos... Obrigada por tudo e por fazer parte da minha vida!

Por fim, preciso agradecer às instituições de fomento (em especial ao CNPq pelas bolsas de doutorado no país e sanduíche), às instituições federais, e a este País, chamado Brasil. Tive a oportunidade de ter estudado sempre em escolas e universidades públicas e ter tido bolsas de estudo para me dedicar exclusivamente às atividades acadêmicas. Esse suporte foi não somente importante, mas fundamental para que chegasse até aqui.

Sumário

Resumo

Os elementos de transposição (TEs) são segmentos de DNA que têm a capacidade de mover-se e replicar-se dentro do genoma. Estão presentes em praticamente todos os organismos, compreendendo uma fração significativa do genoma dos mesmos. Duas classes de TEs são amplamente reconhecidas, os retrotransposons (classe I), que se transpõem por um intermediário de RNA, e os transposons (classe II) que usam DNA como intermediário direto da transposição. A diversidade, complexidade e ubiqüidade dos elementos transponíveis, a ampla variação fenotípica e molecular produzida em seus hospedeiros como conseqüência de sua transposição, assim como a transmissão horizontal da informação genética entre espécies indicam que essas seqüências desempenham uma importante função no processo evolutivo dos genomas, justificando a importância do seu estudo nos diversos organismos. O presente trabalho procurou explorar a história evolutiva de diferentes elementos de transposição em *Drosophila* visando contribuir para o entendimento do processo de co-evolução dessas seqüências com o genoma hospedeiro. Nosso principal foco foi investigar a evolução de retrovírus endógenos de *Drosophila.* Evidenciamos a ocorrência de um grande número de eventos de transmissão horizontal entre espécies. Muitos desses retrovírus podem ainda estar ativos e potencialmente serem agentes infecciosos, o que pode ajudar a explicar o grande número de eventos de transferência horizontal que encontramos. Investigamos também, a distribuição e evolução de uma família de transposons de DNA não autônomos, os quais podem ser considerados elementos do tipo MITEs (Miniature Inverted-repeat Transposable Elements). Nossas análises confirmaram que diferentes processos têm contribuído para a evolução e distribuição dos TEs nos genomas, como transmissão vertical, perda estocástica, polimorfismo ancestral, introgressão e transferência horizontal.

Abstract

Transposable elements (TEs) are segments of DNA that have the ability to move and replicate within the genome. They are present in nearly all organisms, composing a significant fraction of their genomes. Two classes of TEs are widely recognized, the retrotransposons (class I) that transpose through a RNA intermediate and transposons (class II) that use DNA as a direct intermediate of transposition. The diversity, complexity and ubiquity of transposable elements, the extensive phenotypic and molecular variation produced in their hosts as a consequence of its transposition, as well as genetic horizontal transmission between species, indicate that TEs play an important role in evolution of genomes, substantiating the importance their study in different organisms. This study aimed to explore the evolutionary history of different transposable elements in *Drosophila* to contribute to the understanding of the co-evolution of these sequences with the host genome. Our main focus was to investigate the evolution of *Drosophila* endogenous retroviruses and we found several examples of horizontal transfer between species. Some of these retroviruses may still be active and are potentially infectious agents, which help to explain the high number of horizontal transfer events. We also investigate the distribution and evolution of a non-autonomous family of DNA transposons, which can be considered as MITEs (Miniature Inverted-repeat Transposable Elements). Our analyses confirm that several processes have contributed to the evolution and distribution of transposable elements in the genomes, such as vertical transmission, stochastic loss, ancestral polymorphism with independent assortment of copies during speciation, introgression and horizontal transfer.

Capítulo 1

Introdução

Elementos transponíveis (TEs) são seqüências de DNA que podem se mover para novas posições dentro do genoma de uma única célula. TEs compõem uma grande fração do genoma das células eucarióticas e são vistos como poderosos fatores que influenciam a evolução das espécies hospedeiras, tanto a curto como em longo prazo. Essas seqüências móveis foram descobertas por Barbara McClintock, nos anos 40, sendo ela uma das primeiras pesquisadoras a sugerir que os TEs podem desempenhar algum papel regulador nos genomas. McClintock recebeu o Prêmio Nobel de Fisiologia ou Medicina em 1983, 40 anos após a sua descoberta.

Os TEs têm sido encontrados virtualmente em todos os organismos investigados (Wicker et al. 2007). Porém, publicações de alguns genomas eucarióticos unicelulares relatam a ausência de TEs nos mesmos. Incluídos nessa lista estão a alga vermelha *Cyanidioschyzon merolae*, os Apicomplexas: *Babesia bovis*, *Cryptosporidium hominis*, *C. parvum*, *Plasmodium falciparum, P. yoelli* e *Thelieria parva*. No entanto, como esses organismos são distantemente relacionados da maioria dos genomas eucarióticos com seqüências disponíveis, a falta de TEs relatados, em alguns casos, pode refletir uma incapacidade de identificação dos mesmos baseando-se na similaridade com os tipos conhecidos (Phritam 2009).

O número de cópias de um dado TE em um genoma pode variar de poucas a milhares. Cada hospedeiro pode ter muitos tipos diferentes de TEs, os quais podem representar parte considerável de seus genomas. Por exemplo, quase 85% do genoma do milho (Schnable et al. 2009) e 45% do genoma humano (Lander et al. 2001) são constituídos por TEs. Em *Drosophila*, a fração composta por TEs varia nos diferentes genomas. *D. simulans* e *D. grimshawi* apresentam apenas

2,7% de seqüências repetitivas, enquanto essa proporção passa para 25% em *D. ananassae* e 15,57% em *D. willistoni* (*Drosophila* 12 Genomes Consortium 2007).

A abundância e a diversidade dessas seqüências originam uma série de questões sobre a natureza das relações entre TEs e seus hospedeiros, bem como sobre as implicações desses elementos na evolução dos genomas. O que fazem essas seqüências nos genomas? Como surgiram? Como os TEs são mantidos por longos períodos evolutivos? Qual a dinâmica dessas seqüências nos genomas? Como são controlados? Quais são as conseqüências, vantagens e desvantagens, da presença de TEs nos genomas? São os TEs parasitas, egoístas?

1. Classificação dos TEs

Os TEs são divididos em duas grandes classes de acordo com sua estrutura e modo de transposição. Os elementos da classe I são aqueles que se transpõem por um intermediário de RNA, o qual sofre transcrição reversa para a inserção. São os chamados retrotransposons. Os elementos da classe II usam DNA como intermediário direto da transposição, e são chamados transposons (Finnegan 1989). Uma classificação mais detalhada considera a estrutura geral dos elementos, a existência de domínios, motivos e assinaturas protéicas semelhantes e o grau de similaridade de seqüências de nucleotídeos ou aminoácidos para definir subclasses, superfamílias, famílias e subfamílias (Capy et al. 1998).

Wicker et al. (2007) propuseram um sistema de classificação hierárquica unificado, projetado com base no mecanismo de transposição, na semelhança de seqüência e relações estruturais, que podem ser facilmente aplicadas por não-especialistas. Esse sistema inclui os níveis de classe, subclasse, ordem, superfamília, família e subfamília.

Segundo Wicker et al. (2007) os retrotransposons são divididos em cinco ordens com base na organização e filogenia da transcriptase reversa: LTR-retrotransposons, DIRS, PLP (Penelope*)*, LINEs e SINEs. Essa divisão foi importante para a classificação dos elementos *DIRS-like* e *Penelope-like*, os quais apresentam características distintas dos retrotransposons que eram tradicionalmente conhecidos. Cada uma dessas ordens é subdividida em superfamílias. A ordem dos retrotransposons com LTRs (*long terminal repeats* - longas repetições terminais), foco principal dessa tese, será discutida em um tópico a parte.

Os elementos de classe II são divididos em duas subclasses de acordo com o número de fitas de DNA que são cortadas durante a transposição. Subclasse 1 são mobilizados pelo mecanismo conhecido como corta e cola, clivando ambas as fitas durante a transposição. Subclasse 2 inclui elementos que clivam apenas uma das fitas de DNA durante o processo de transposição. A principal ordem da subclasse I, ordem TIR (*terminal inverted repeat* - repetições terminais invertidas), compreende os transposons clássicos, caracterizados pela presença de TIRs. Nesses elementos a transposição é mediada pela enzima transposase que reconhece as TIRs e corta ambas as fitas. Uma característica compartilhada com a maioria dos TEs é a presença de TSDs (*target site duplication*) que são duplicações no sítio alvo durante o processo de transposição. Elementos TIR são divididos em 9 superfamílias de acordo com a seqüência das TIRs e tamanho das TSDs (Wicker et al. 2007). A superfamília *hAT* será também discutida a parte.

No sistema de classificação hierárquica proposto por Wicker et al. (2007) um TE pode ser classificado em poucas etapas e parece bastante útil para auxiliar a identificação e anotação de TEs nos genomas, porém apresenta algumas deficiências e a comunidade científica apresenta certa resistência ainda quanto a essa classificação. Possivelmente, nos próximos anos será proposto um novo sistema, pois existe um comitê internacional para classificação dos elementos de transposição, que foi instituído durante o "1st International Conference/Workshop Genomic Impact of Eukaryotic Transposable Elements" que aconteceu em 2006, nos Estados Unidos.

2. LTR-Retrotransposons

Uma grande fração dos genomas eucarióticos é composta por retrotransposons, os quais se transpõem por um intermediário de RNA. Retrotransposons com LTRs são evolutivamente relacionados aos retrovírus de vertebrados, com os quais compartilham similaridades estruturais e funcionais.

Os retrovírus compreendem uma grande e diversificada família (Retroviridae) de vírus de RNA envelopados. Os retrovírus típicos são agentes infecciosos, como por exemplo, o HIV, e possuem um genoma de RNA. Uma vez dentro da célula esse genoma é copiado em uma molécula de DNA para ser inserida no genoma da célula infectada. Essa integração é uma etapa essencial do ciclo de vida dos retrovírus. A forma integrada no genoma é referida como provírus. Um provírus típico (Figura1) consiste de longas repetições terminais flanqueando uma região central que contém três fases abertas de leitura (ORFs – *open reading frame*) que constituem os genes *gag*, *pol* e *env*, os quais são requeridos para replicação viral e infecciosidade (Coffin et al. 1997).

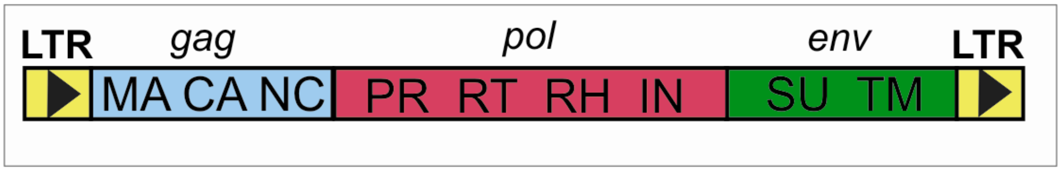


Figura 1: Organização estrutural básica de um provírus. Duas LTRs flanqueando os genes *gag*, *pol* e *env.* MA – Matriz; CA – Capsídeo; NC – Nucleocapsídeo; PR – Protease; RT – Transcriptase Reversa; RH – Ribonuclease H; IN – Integrase; SU – Glicoproteína de Superfície; TM – Peptídeo Transmembrana.

O gene *gag* codifica uma proteína estrutural interna do vírus, a proteína Gag (*group-specific antigen*), a qual é proteoliticamente processada em três proteínas maduras: Matriz (MA), Capsídeo (CA) e Nucleocapsídeo (NC), correspondendo aos componentes da partícula viral (VLP – *virus-like particle*). O gene *pol* (*polymerase*) codifica todas as proteínas requeridas para transposição, incluindo Protease (PR), Transcriptase Reversa (RT), Ribonuclease H (RH) e Integrase (IN). O gene *env* (*envelope*) codifica as proteínas de superfície (SU) e transmembrana (TM) do vírion, as quais formam um complexo que interage especificamente com receptores celulares, que levam à fusão do vírion com a membrana da célula. O envelope viral é formado pela bicamada lipídica derivada da célula na qual as proteínas Env são inseridas (Boeke e Stoye 1997).

Os retrovírus, usualmente, somente infectam células somáticas e conseqüentemente, os genes retrovirais não são passados para a progênie do hospedeiro. Alguns tipos de retrovírus, no entanto, podem ocasionalmente infectar a linhagem germinativa. A progênie resultante das células germinativas infectadas irá carregar o provírus como parte de seu genoma e esses retrovírus serão subseqüentemente transmitidos verticalmente de uma geração para outra. Esses retrovírus são referidos como retrovírus endógenos (ERVs - *endogenous retroviruses*) e podem persistir no genoma de seus hospedeiros por longos períodos. Eles são geralmente silenciados transcricionalmente e muitas vezes são defectivos e incapazes de formar vírus infeccioso (Coffin et al. 1997; Terzian et al. 2001).

Retrovírus e retrotransposons compartilham um mesmo mecanismo de replicação com a diferença que retrotransposons típicos não possuem o gene *env* e dessa forma, não fazem partículas infecciosas, que saem de uma célula para infectar outras. A figura 2 sumariza o ciclo de vida dos retrotransposons e retrovírus. Essa distinção entre retrotransposon e retrovírus poderia ser simples no passado. No entanto, a presença de um gene *env-like* tem sido identificada em vários representantes de LTR-retrotransposons (Malik et al. 2000), como será discutido mais adiante.

Em relação à classificação dos LTR-retrotransposons, Capy et al. (1998) sugerem que sejam subdivididos em dois grupos, os quais diferem na ordem dos domínios enzimáticos codificados pelo gene *pol*: *Ty1/copia* e *Ty3/gypsy*. No primeiro grupo a ordem dos domínios é 5’ PR-IN-RT-RH 3’ e no segundo a ordem é 5’ PR-RT-RH-IN 3’. A classificação apresentada por Wicker et al. (2007) é um pouco mais específica e divide os LTR-retrotransposons em cinco superfamílias: *Copia, Gypsy, Bel-Pao, Retrovirus* e *ERV*. *Copia* e *Gypsy* correspondem aos grupos *Ty1/copia* e *Ty3/gypsy* anteriormente conhecidos e a superfamília *Bel-Pao* é estruturalmente similar aos elementos *Gypsy e Copia*, porém forma um clado distinto em filogenias baseadas na RT. Interessantemente, os retrovírus de vertebrados foram também incluídos nessa classificação em duas superfamílias, *Retrovirus,* que compreendem as formas infecciosas e *ERV*, que são as formas endógenas.

Sem levar em consideração os sistemas de classificação, mas apenas a filogenia da RT obtida para retroelementos com LTRs, eles podem ser divididos em quatro principais clados (Figura 3): os LTR-retrotransposons *Ty1-copia*, *Ty3-gypsy* e *Bel-Pao* e os retrovírus de vertebrados, Retroviridae (Eickbush e Jamburuthugoda, 2008). Filogenias da RT têm sugerido que os retrovírus de vertebrados são derivados de retrotransposons com LTRs pela aquisição de um gene *env* (Xiong e Eickbush, 1990; Malik et al. 2000). Essa transição de retrotransposon não viral para retrovírus parece ter ocorrido diversas vezes na história evolutiva dos retroelementos. Eventos de aquisições independentes explicam a presença de gene *env* em vários representantes dos grupos *Ty1-copia*, *Ty3-gypsy* e *Bel-Pao* (Malik et al. 2000). Dessa forma vários retrotransposons são também chamados de retrovírus endógenos.

Não está claro se os genes *env* de retrovírus de vertebrados representam um evento ou múltiplos eventos de aquisição. O grupo *Ty1/copia* possui um exemplo de aquisição de gene *env* no elemento *SIRE1* de soja, *Glycine max*. O grupo *Bel-Pao* contém três eventos de aquisição do gene *env: Cer7* de *C*. *elegans*, *Tas* de *Ascaris lumbricoides* e *Roo-like* em *Drosophila.* O clado dos elementos *Ty3-gypsy* possui pelo menos três prováveis aquisições independentes do gene *env*, *Athila* em *Arabidopsis*, *Osvaldo* e *gypsy-like* em *Drosophila* (Malik et al. 2000; Malik e Henikoff 2005).

O oportunismo, demonstrado por LTR-retrotransposons, em adquirir um gene *env* de outro agente infeccioso, apresenta um modelo geral em que, potencialmente, qualquer LTR-retrotransposon pode se tornar um vírus (Malik et al. 2000).

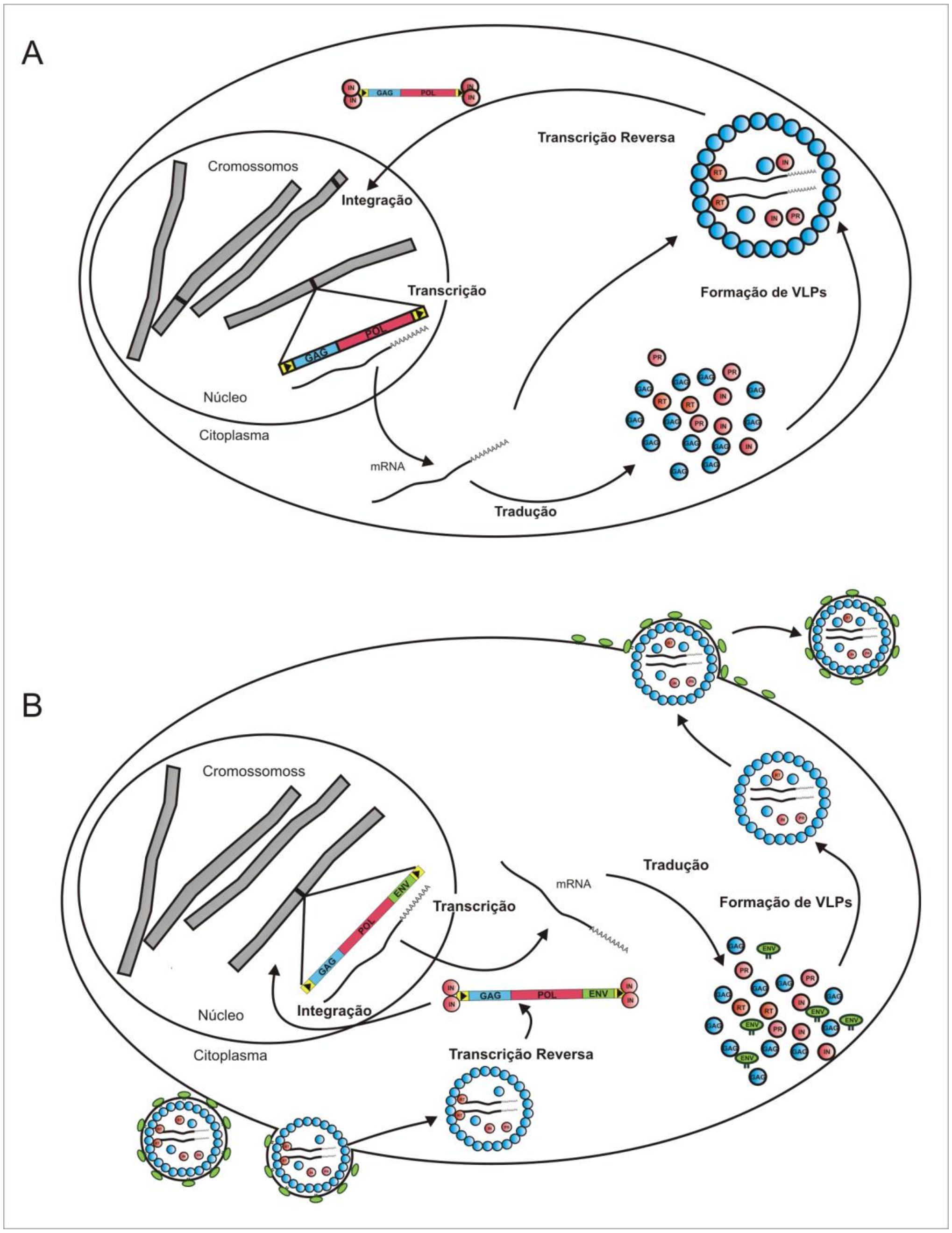


Figura 2: Ciclo de vida de LTR-retrotransposons (A) e retrovírus (B). (A) Primeiramente, o retrotransposon é transcrito e o RNA é então traduzido no citoplasma para gerar as proteínas que formam VLPs. Normalmente, duas moléculas do RNA do retrotransposon são empacotadas em uma VLP, onde o RNA é posteriormente transcrito reversamente em uma cópia de DNA, a qual é integrada no genoma hospedeiro, acrescentando outra cópia do retrotransposon ao genoma. (B) Após a ligação do vírus aos receptores na superfície da célula, ocorre fusão das membranas viral e celular, e o vírus entra na célula. A transcrição reversa do RNA viral ocorre dentro do complexo de nucleoproteínas. O DNA viral entra no núcleo e integra-se no genoma para formar um provírus. O provírus é transcrito, traduzido e novas partículas virais são montadas na membrana plasmática.

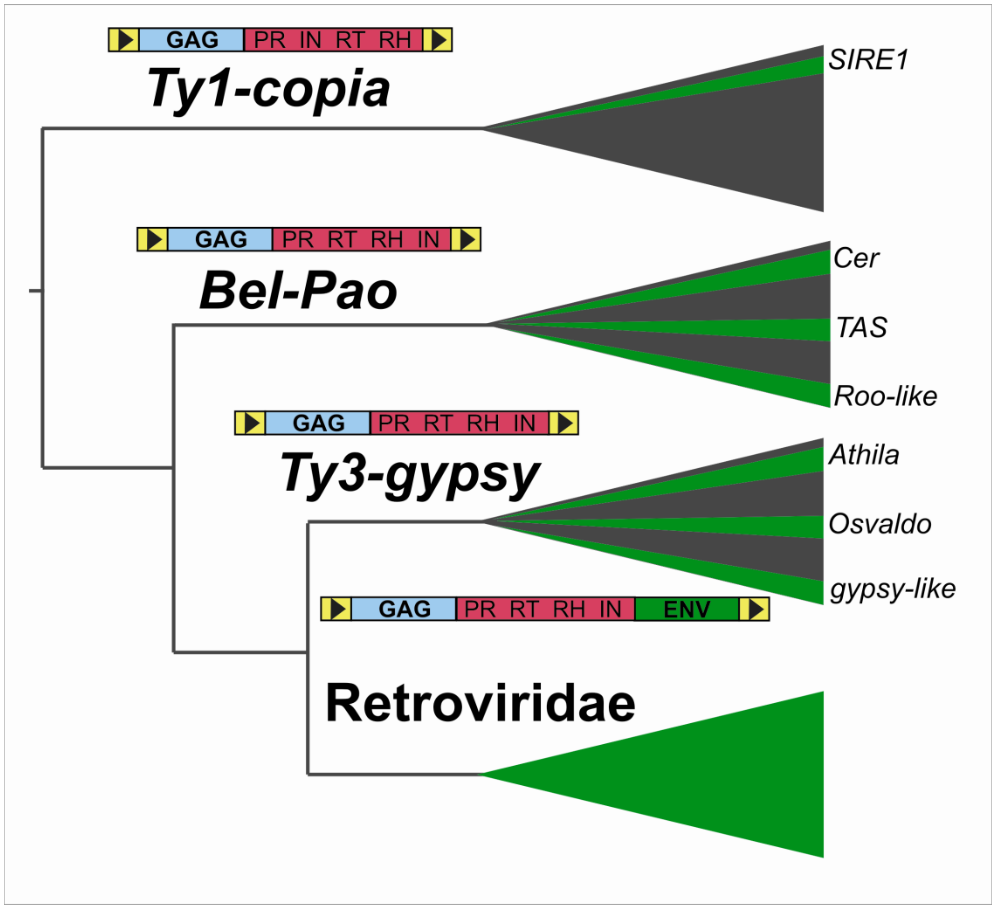


Figura 3: Relações filogenéticas dos retrotroelementos com LTRs e organização básica de suas seqüências. Eles são divididos em quatro grupos, os quais são bastante diversos e foram representados pelos triângulos. As regiões em verde representam linhagens evolutivas que possuem o gene *env*. Adaptada de Malik et al. (2000) e Eickbush e Jamburuthugoda (2008).

Levando em consideração as relações filogenéticas e a presença do gene *env*, cada um dos quatro clados citados pode ser subdivido em diversas linhagens evolutivas. Uma das linhagens evolutivas do grupo *Ty3-gypsy* compreende os elementos por vezes referidos como *gypsy-like* presentes em insetos. Diversas famílias de elementos *gypsy-like* são descritas em *Drosophila* e outros insetos (Bowen e McDonald 2001; Kaminker et al. 2002; Kapitonov e Jurka 2003). Malik et al. (2000) identificaram blocos conservados de aminoácidos do gene *env* dos elementos *gypsy-like* com uma similaridade significativa com proteínas de fusão (FP – fusion protein) de baculovírus de insetos. Os baculovírus formam uma grande e diversa família (Baculoviridae) de vírus patogênicos de insetos, com genoma de DNA fita dupla, circular e supertorcido. As FPs de baculovírus são proteínas *env-like*, que causam a fusão do vírion e da membrana celular.

O elemento *gypsy* é um dos retrotransposons de *Drosophila* mais estudados e possui propriedades infecciosas comprovadas (Kim et al. 1994; Song et al. 1994). Várias outras famílias de retrotransposons potencialmente retêm essa capacidade infecciosa devido à presença de gene *env* funcional. Malik et al. (2000) sugerem que a presença do gene *env* nos retroelementos poderia aumentar a probabilidade de eventos de transferência horizontal. De fato, numerosos casos de transferência horizontal (TH) de retrotransposons *gypsy-like* têm sido descritos nos últimos anos (Terzian et al. 2000; Herédia et al. 2004; Sánchez-Gracia et al. 2005; Ludwig e Loreto 2007; capítulo II, Ludwig et al. 2008; Vidal et al. 2009), vários desses por nosso grupo de pesquisa.

Um evento independente de aquisição de gene *env* de baculovírus aconteceu em uma linhagem de retrotransposons de *Drosophila* do grupo *Bel-Pao* (Malik e Henikoff 2005)*,* a qual chamamos *roo-like*. Essa linhagem apresenta apenas as famílias *roo*, *rooA* e *Kanga* descritas até o momento (de la Chaux e Wagner 2009; Malik e Henikoff 2005). O capítulo 5 apresenta uma análise evolutiva do retrotransposon *Kanga*.

3. Superfamília *hAT* de transposons de DNA

A superfamília *hAT* de transposons é composta por elementos de DNA presentes nos mais diversos organismos como fungos, nematódeos, peixes, insetos, plantas e humanos (Kempken e Windhofer 2001). O nome da superfamília *hAT* é devido a três elementos que a compõem: *hobo*, descrito originalmente em *D. melanogaster* (Calvi et al. 1991), *Activator* (*Ac*) de *Zea mays* (McClintock 1947) e *Tam3* de *Antirrhinum majus* (Hehl et al. 1991).

De acordo com a classificação de TEs sugerida por Wicker et al. (2007) elementos da superfamília *hAT* compartilham algumas características como: (1) TSDs de 8 pb, (2) TIRs relativamente curtas (entre 5 e 27 pb), e (3) geralmente possuem menos de 4 kb de tamanho. Membros da superfamília *hAT* podem ainda ser subdivididos em famílias que são definidas pelo grau de similaridade em suas seqüências de DNA.

Várias famílias de elementos *hAT* foram caracterizadas em insetos, como *hobo* em *D. melanogaster* (Calvi et al. 1991) *Hermes* em *Musca domestica* (Warren et al. 1994), *Hermit* em *Lucilia cuprina* (Coates et al. 1996), *Homer* em *Bactrocera tryoni* (Pinkerton et al. 1999), *hopper* em *B. dorsalis* (Handler e Gomez 1997) e *Herves* em *Anopheles gambiae* (Arensburger et al. 2005). Recentemente, novos elementos *hAT* foram identificadas por Ortiz e Loreto (2009) através de buscas *in silico* no 12 genomas disponíveis de *Drosophila*.

Esta ampla distribuição de elementos *hAT* parece ser devida à origem muito antiga desta superfamília, provavelmente precedendo a separação de plantas, fungos e animais (Rubin et al. 2001). No entanto, alguns casos de TH podem explicar a distribuição destes elementos e as relações entre seqüências *hAT* em alguns grupos de espécies. Eventos de TH parecem fazer parte da história evolutiva de alguns membros da superfamília *hAT* como, por exemplo, *Herves*, transmitido de um doador desconhecido para o genoma de *A. gambiae* (Subramanian et al. 2007); *Tol2,* o qual invadiu duas espécies de peixes do gênero *Oryzias* (Koga et al. 2000) e *Myotis-hAT1* passado para o genoma de morcegos do gênero *Myotis* a partir de uma fonte desconhecida (Ray et al. 2007).

No gênero *Drosophila*, o elemento *hobo* é um dos TEs da superfamília *hAT* mais estudados e aparentemente está envolvido em eventos de TH recentes entre membros do grupo *melanogaster* (Boussy e Daniels 1991; Pascual e Periquet 1991; Simmons 1992). Recentemente, nosso grupo identificou dois elementos da superfamília *hAT*, *harrow e hosimary*, envolvidos em TH (Mota et al. 2010; Deprá et al. 2010).

Alguns trabalhos em *Drosophila* têm mostrado a presença de pequenos TEs derivados de elementos *hAT* (Holyoakee e Kidwell 2003; Ortiz e Loreto 2008; Ortiz et al. 2010). Esses elementos podem ser considerados MITEs, que serão discutidos no próximo tópico.

4. *Miniature Inverted-repeat Transposable Elements* - MITEs

Elementos de classe I e de classe II possuem elementos autônomos e não-autônomos. Somente os elementos autônomos possuem todas as seqüências que codificam as proteínas essenciais para sua propagação. No entanto, elementos não-autônomos são mobilizados pelas atividades enzimáticas providas por elementos autônomos (Kidwell e Lisch 2001).

MITEs formam um grupo heterogêneo de pequenos elementos de DNA não-autônomos, que são flanqueados por TIRs e são freqüentemente encontrados dentro ou próximo a genes. A designação de elementos como MITEs não reflete uma comum origem dessas seqüências, porém essa designação é bastante útil para referir-se a esses TEs com características particulares comuns em diferentes famílias (Wicker et al. 2007).

MITEs são amplamente distribuídos em genomas eucarióticos, onde eles podem acumular um alto número de cópias apesar da falta de capacidade codificadora. Eles têm sido extensivamente estudados em plantas, em particular no genoma do arroz (Gonzáles e Petrov 2009).

Entretanto, pouco é conhecido sobre a origem e amplificação desses TEs. Alguns MITEs são claramente derivados de alguma família autônoma de TEs, como *mPing* derivada de *Ping* em arroz (Jiang et al. 2003). No entanto, a grande maioria dos MITEs descritos não está diretamente relacionada a uma família específica de TEs. O modelo mais lógico para explicar a transposição dessas seqüências é que eles devem utilizar a transposase de elementos mais distantemente relacionados (Yang et al. 2009).

O capítulo 4 dessa tese apresenta um estudo sobre um MITE de *Drosophila*, o elemento *Mar.* Esse elemento foi caracterizado em *D. willistoni,* possui 610 pb, TSDs de 8 pb, TIRs de 11 pb e não apresenta capacidade codificadora. Esse elemento foi classificado como um MITE da superfamília *hAT* com base no tamanho das TIRs e TSDs.

5. Evolução dos Elementos de Transposição

Para entender a evolução de um TE em particular dentro de um gênero é necessário analisar a distribuição e a conservação do elemento nas espécies deste gênero. Quando TEs são transmitidos verticalmente sua história filogenética deve refletir a história evolutiva dos seus hospedeiros. Desse modo, a filogenia dos TEs é freqüentemente comparada com a filogenia das espécies hospedeiras (Silva et al. 2004).

Segundo Silva et al. (2004) três tipos de distorções da filogenia esperada de um TE são usualmente utilizadas para detectar transmissão horizontal:

(1) Detecção de elementos com alto grau de similaridade de seqüência em táxons não relacionados. Neste caso, a divergência entre as seqüências de TE é muito menor do que a divergência entre genes nucleares das suas respectivas espécies hospedeiras.

(2) Detecção de diferenças topológicas entre a filogenia do TE e das espécies hospedeiras.

(3) Distribuição descontínua dos elementos entre táxons proximamente relacionados. A presença de um TE em uma linhagem e a ausência em uma linhagem irmã, resulta na ausência de um ou mais ramos na filogenia do TE.

Dessa forma, muitas filogenias e distribuições específicas de TEs poderiam ser explicadas por transmissão horizontal entre espécies. Entretanto, explicações alternativas precisam ser testadas, como polimorfismo ancestral com independente distribuição de cópias nas espécies descendentes, taxas diferentes de substituição em TEs nas diferentes espécies e perda estocástica de TEs em alguns táxons.

Numerosos casos de transmissão horizontal de TEs entre espécies de *Drosophila* têm sido reportados envolvendo tanto elementos de classe II, como *P* (Daniels et al. 1990b; Silva e Kidwell 2000; Loreto et al. 2001), *mariner* (Maruyama e Hartl 1991), *hobo* (Daniels et al. 1990a), como também envolvendo elementos da classe I, destacando-se o retroelemento *gypsy* (Herédia et al. 2004).

A maioria dos casos de TH descritos envolve espécies distantemente relacionadas, apresentando óbvias diferenças entre divergências do TE e de genes do hospedeiro. No entanto, análises de cinco espécies proximamente relacionadas do subgrupo *melanogaster,* as quais possuem genoma disponível, têm sugerido ocorrência de TH entre essas espécies. Nesses casos, o TE apresenta uma divergência menor que genes nucleares. Porém, pela proximidade das espécies os genes nucleares também apresentam uma pequena divergência. Dessa forma, cresceu a preocupação do nosso grupo em desenvolver metodologias e utilizar testes estatísticos para estabelecer eventos de TH, um dos objetivos do capítulo 2. Nesse capítulo é apresentado um trabalho onde utilizamos uma abordagem diferente para a inferência de TH. Nós incluímos um teste estatístico, e empregamos comparações de valores substituição sinônima (dS) dos TEs com os encontrados para genes que apresentam um índice similar de uso de códons, já que está característica tem implicações nos valores de dS.

6. Implicações Evolutivas dos TEs

Muitos estudos têm revelado o impacto dos TEs na função e estrutura do genoma e na evolução e adaptação das populações (revisão em Kidwell e Lisch 2001; Bowen e Jordan 2002; Biemont e Vieira 2006; Wessler 2004; Volff 2006; Pritham 2009; Oliver e Greene 2009).

Segundo Oliver e Greene (2009) os TEs podem gerar novidades evolutivas de duas maneiras diferentes: (1) ativamente, via inserção *de novo* contribuindo diretamente com seqüencias codificadoras ou alteração de regulação, ou ainda por retrotransposição de transcritos gerando duplicação gênica; e (2) passivamente, promovendo recombinação ectópica que resulta em rearranjos cromossômicos, duplicações ou deleções (Cáceres et al. 1999; Casals et al. 2003; Delprat et al. 2009).

Muitos exemplos de domesticação molecular têm sido descritos, onde os TEs são recrutados para desempenhar uma função no hospedeiro (McDonald 1993; Britten 1996; Nekrutenko e Li 2000; Kidwell e Lisch 2001; Bundock e Hooykaas 2005; Hammer et al. 2005).

Em *Drosophila* foi identificado um evento de domesticação, relativamente recente, de um gene *env* retroviral (Malik e Henikoff 2005). Esse gene representa o gene *env* domesticado da linhagem *Kanga*, um novo clado de retrovírus endógenos *Bel-Pao* da linhagem *roo-like*. *Iris* é preferencialmente expresso em adultos, tanto em fêmeas quanto em machos, sugerindo que esse gene foi domesticado para desempenhar alguma função em moscas adultas. A função do gene *Iris* ainda não foi esclarecida, mas os autores propõem que *Iris* tenha sido recrutado como um gene hospedeiro especificamente para defender adultos contra recorrentes invasões por retrovírus ou baculovírus, os quais compartilham um *env* homólogo. O capítulo 5 dessa tese apresenta um estudo evolutivo do retrotransposon *Kanga*, que deu origem ao gene *Iris*.

7. Regulação dos TEs em *Drosophila*

Quando os TEs apresentam alta taxa de transposição, eles podem ser prejudiciais ao hospedeiro. Portanto, mecanismos que silenciam TEs foram selecionados para estabilizar os genomas (Kidwell e Lisch 2001). Essa seção é uma tentativa de sumarizar os conhecimentos sobre a regulação de TEs mediada por pequenos RNAs em *Drosophila*. Os componentes chave desse mecanismo são as proteínas Piwi e os pequenos RNAs que interagem com essas proteínas (piRNAs – *Piwi-interacting RNAs*).

Primeiramente, o silenciamento de genes mediado por distintos pequenos RNAs de 20-30 nucleotídeos de comprimento é coletivamente chamado de RNA de interferência ou silenciamento de RNA. Basicamente, neste mecanismo, os pequenos RNAs formam um complexo com as proteínas da família Argonauta, chamado complexo RISC (*RNA-induced silencing complex*). Nesse complexo, as proteínas Argonauta e pequenos RNAs desempenham papéis diferentes: os pequenos RNAs guiam as proteínas Argonauta aos seus alvos, enquanto as proteínas Argonauta exercem atividades enzimáticas para inibir a expressão gênica, principalmente por clivagem de transcritos ou bloqueando a síntese protéica (Liu et al. 2004). Diferentes tipos de pequenos RNAs (miRNA, siRNA, piRNA) são encontrados nas células, os quais possuem diferentes tamanhos, são gerados por mecanismos distintos e possuem diferentes mecanismos de atuação e funções (Obbard et al. 2009).

Em animais existem dois grupos de proteinas Argonauta. O primeiro grupo contém proteínas que agem com miRNAs e siRNAs e o segundo contém as proteínas da subfamília Piwi que interagem com os piRNAs. Em *Drosophila* o clado *Piwi* consiste de três membros: *Piwi*, *Aubergine* (*Aub*) e *Ago3* (Carmell et al. 2002).

Em *Drosophila*, diferentes lócus heterocromáticos são a fonte primária de piRNAs antisenso, os quais, então, silenciam um grande número de TEs que estão dispersos no genoma e são ativos na linhagem germinativa (Brennecke et al. 2007). Esses autores propuseram um ciclo de amplificação de piRNAs, denominado *ping-pong*, o qual parece ser um dos principais mecanismos que regulam a atividade de transposons. Nesse mecanismo, piRNAs antisenso associam-se a Piwi ou Aub e marcam como alvo os transcritos de elementos funcionais presentes na eucromatina e após a clivagem geram piRNAs senso. Os piRNAs senso associam-se então a proteína Ago3 e podem ser usados para produzir adicionais piRNAs antisenso direcionando a clivagem de transcritos antisenso sintetizados a partir dos *clusters* heterocromáticos (Brennecke et al. 2007).

O trabalho de Brennecke et al. (2007) resolveu um grande "mistério" que envolvia a regulação do retrotransposon *gypsy* pelo lócus *flamenco.* Elegantes estudos genéticos mostraram que o lócus *flamenco* é um regulador chave para controlar a atividade de *gypsy* e foi mapeado na região heterocromatica pericentromérica do cromossomo X de *D. melanogaster* (Pelisson et al. 1994; Prud'homme et al. 1995). No entanto, pelas dificuldades de acessar a seqüência desse lócus heterocromático e pela presença de várias seqüências repetitivas, não se sabia de que forma *flamenco* silenciava os retroelementos. Sarot et al. (2004) sugeriam que o silenciamento de RNA deveria ser o processo envolvido na repressão de *gypsy* nos ovários contendo os alelos *flamenco* e *piwi* funcionais, porém o papel de *flamenco* nesse processo ainda não estava esclarecido. Com uso do seqüenciamento de nova geração Brennecke et al. (2007) seqüenciaram mais de 50000 *reads* referentes à pequenos RNAs (de 23 a 29 pb) associados a Piwi, Ago3 e Aub. Eles mostraram que os piRNAs são predominantemente seqüências de transposons antisenso (para Piwi e Aub) e senso (para Ago3). Esses piRNAs formam *clusters* que pareiam em múltiplos locais nos cromossomos, a partir de onde eles são transcritos. Os autores mostraram que o lócus *flamenco* é precisamente um desses clusters responsáveis pela geração de piRNAs.

Recentemente, Malone et al. (2009) identificaram duas vias de regulação por piRNA com distintos componentes nas células germinativas e somáticas do ovário. Já havia sido demonstrado que muitos TEs são expressos nas células germinativas, onde a atividade de transposição pode criar uma substancial carga mutacional que acumula ao passar das gerações. Exemplos em *Drosophila* incluem *TAHRE, TART, HetA*, *copia.* Outros TEs são exclusivamente ou adicionalmente expressos nas células somáticas dos ovários, como os retrotransposons *gypsy-like*, *gypsy, ZAM* e *idefix.* Malone et al. (2009) mostraram que nas células foliculares somáticas, a proteína Piwi age unicamente com o *cluster* de piRNA *flamenco* para realizar o silenciamento de retrovírus endógenos que podem se propagar infectando as células germinativas. O mecanismo envolvido é distinto do *ping-pong* anteriormente sugerido. Já nas células germinativas, diferentes *clusters* de piRNAs colaboram com as três proteínas da subfamília Piwi, pelo mecanismo *ping-pong,* para controlar um abrangente espectro de TEs.

Existe também um segundo mecanismo de RNAi anti-TE em *Drosophila,* o qual utiliza siRNAs endógenos (21 nucleotídeos) processados, pelas enzimas Ago-2 e Dicer-2, gerados a partir do transcrito do TE endógeno*.* Esse mecanismo ocorre tanto nas células somáticas como germinativas do ovário (Chung et al. 2008).

Objetivos

A presente tese teve como objetivo geral explorar a história evolutiva de elementos de transposição em *Drosophila*, visando contribuir para o entendimento das relações entre os elementos de transposição e espécies hospedeiras. Os objetivos específicos foram:

Capítulo 2:

(1) Estudo e aplicação de novas metodologias para inferência de casos de transmissão horizontal.

(2) Acessar a história evolutiva dos retrovírus endógenos de *Drosophila gypsy, gypsy2, gypsy3, gypsy4* e *gypsy6.*

Capítulo 3:

(1) Analisar a distribuição, conservação e evolução do retrovírus endógeno *nik* em *Drosophila*.

(2) Acessar informações sobre a regulação de *nik* nos genomas.

Capítulo 4:

(1) Investigar a distribuição, conservação e evolução do elemento *Mar*, em *Drosophila.*

(2) Identificar elementos autônomos que podem ser responsáveis pela mobilização de *Mar*.

Capítulo 5:

(1) Contribuir para o entendimento da origem do gene domesticado *Iris* em *Drosophila*.

(2) Caracterização do retrovírus endógeno *Kanga* e análise da sua distribuição, conservação e evolução em *Drosophila.*

Capítulo 2

Multiple invasions of *Errantivirus* in the genus *Drosophila**

Adriana Ludwig[1], Vera L. da S. Valente[1] and Elgion L. S. Loreto[1,2]

1 *Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil;*

2 *Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), Campus Universitário, Santa Maria, Rio Grande do Sul, Brazil*

* A parte inicial desse trabalho, referente às buscas *in silico* e análises preliminares, foi apresentado na dissertação de mestrado de Adriana Ludwig

Material suplementar do artigo encontra-se nos anexos da tese.

Artigo publicado na revista Insect Molecular Biology

Insect Molecular Biology (2008) **17**(2), 113–124 doi: 10.1111/j.1365-2583.2007.00787.x

Multiple invasions of *Errantivirus* in the genus *Drosophila*

A. Ludwig*, V. L. da S. Valente* and E. L. S. Loreto*†

**Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil; and †Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), Campus Universitário, Santa Maria, Rio Grande do Sul, Brazil*

Abstract

Aiming to contribute to the knowledge of the evolutionary history of *Errantivirus*, a phylogenetic analysis of the *env* gene sequences of *Errantivirus gypsy*, *gtwin*, *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* was carried out in 33 Drosophilidae species. Most sequences were obtained from *in silico* searches in the *Drosophila* genomes. The complex evolutionary pattern reported by other authors for the *gypsy* retroelement was also observed in the present study, including vertical transmission, ancestral polymorphism, stochastic loss and horizontal transfer. Moreover, the elements *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* were shown to have followed an evolutionary model that is similar to *gypsy*. Fifteen new possible cases of horizontal transfer were suggested. The infectious potential of these elements may help elucidate the evolutionary scenario described in the present study.

Keywords: horizontal transfer, *gypsy*, *gypsy2*, *gypsy3*, *gypsy4*, *gypsy6*, *gtwin*, complex evolution, transposable elements, *env* gene.

Introduction

The errantiviruses are insect long terminal repeat (LTR) retrotransposons, characterized by the presence of the *env* gene, in addition to the *gag* and *pol* genes. These elements show considerable structural similarity to vertebrate retroviruses and are named insect endogenous retroviruses, which belong to the Metaviridae family, according to the International Committee on Taxonomy of Viruses (ICTV) (Boeke *et al*., 1999).

The most commonly studied errantivirus is the *gypsy* retroelement of *Drosophila melanogaster*, which shows infectious properties under specific conditions (Kim *et al*., 1994; Song *et al*., 1994). Sequences homologous to the *gypsy* retroelement are widely distributed across the genus *Drosophila* (Stacey *et al*., 1986; Terzian *et al*., 2000; Vázquez-Manrique *et al*., 2000; Herédia *et al*., 2004, 2007) and related elements have been identified in the *D. melanogaster* genome sequence (Jurka, 2000; Bowen & McDonald, 2001; Kaminker *et al*., 2002; Kapitonov & Jurka, 2003).

Like any typical endogenous retrovirus, *gypsy* is vertically transmitted as part of the host genome, although horizontal transfer (HT) events of this retroelement have been suggested for several *Drosophila* species (Alberola & de Frutos, 1996; Terzian *et al*., 2000; Vázquez-Manrique *et al*., 2000; Herédia *et al*., 2004). It has been hypothesized that this ability of the *gypsy* element to undergo HT is related to the presence of the functional *env* gene, which in turn is responsible for the infectious properties of retroviruses (Mejlumian *et al*., 2002; Herédia *et al*., 2004).

This study analyses an approximately 500-bp region of the *env* gene of the *gypsy* and related retroelements in several Drosophilidae species. New sequences for this region were obtained by cloning and sequencing of the *gypsy* element in the *D. flavopilosa* group species and by *in silico* searches of sequences homologous to *gypsy*, *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* from the *Drosophila* genome sequences currently available. The sequences obtained were analysed together with those deposited in GenBank. Twenty-seven probable HT cases were identified, of which 12 have previously been described (Alberola & de Frutos, 1996; Vázquez-Manrique *et al*., 2000; Herédia *et al*., 2004; Ludwig & Loreto, 2007). The results of the present study agree with the findings by Herédia *et al.* (2004), which presented a complex evolutionary model for *gypsy*. It was observed that *gypsy2*, *gypsy4* and *gypsy6* seem to be involved in HT events.

Results

gypsy *in the* D. flavopilosa *group*

The *D. flavopilosa* group is formed by species that present a considerably restricted ecology. These species use

Received 24 August 2007; accepted after revision 3 December 2007; first published online 15 February 2008. Correspondence: Elgion Lúcio Silva Loreto, Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), Campus Universitário, Santa Maria, Rio Grande do Sul, Brazil, 97105-900. Tel.: +55 55 3220 8912; fax: +55 55 3220 8912; e-mail: elgionl@gmail.com

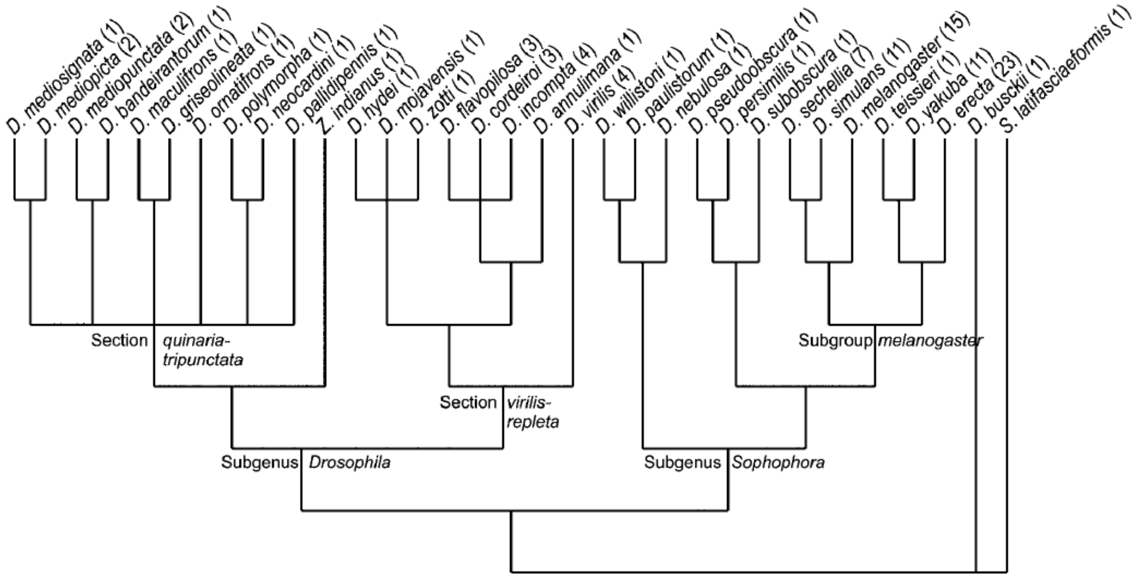


Figure 1. Representation of the evolutionary relationships of the 33 Drosophilidae species analysed in this study based on O'Grady & Kidwell (2002), Lewis *et al.* (2005) and Robe *et al.* (2005). For each species, the number of sequences used is shown in parentheses.

flowers of the genus *Cestrum* (Solanaceae) as exclusive oviposition and larval development sites, with insubstantial habitat overlapping with other *Drosophila* species (Wheeler *et al.*, 1962; Brncic, 1978). An approximately 500-bp fragment of the *gypsy env* gene was amplified in three species of the group: *D. flavopilosa*, *D. incompta* and *D. cordeiroi*. Amplicons were cloned and sequenced. Accession numbers of the sequences in GenBank are EU068744 to EU068753.

Nucleotide sequences of the *gypsy* clones of *D. cordeiroi* and *D. incompta* are similar (between 2.8 and 6.3% of divergence); nevertheless, *D. flavopilosa gypsy* clones presented on average 30% divergence compared to the *D. incompta* and *D. cordeiroi* sequences. The cordeiroiC, cordeiroiD and incomptaM sequences potentially encode for the *env* gene region analysed, but the three *D. flavopilosa* sequences present several indels and nonsense mutations, which suggests old inactivity of these sequences.

Search in the genomes

An *in silico* search was conducted for sequences homologous to the *env* gene of the *gypsy*, *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* retroelements in the genomes of *D. grimshawi*, *D. virilis*, *D. mojavensis*, *D. willistoni*, *D. persimilis*, *D. pseudoobscura*, *D. ananassae*, *D. erecta*, *D. yakuba*, *D. melanogaster*, *D. simulans* and *D. sechellia*. No homologous sequences were found in *D. grimshawi* and *D. ananassae*. For all other species, at least one sequence was found. The absence of the element in *D. grimshawi* and *D. ananassae* could be a result of: (1) the rapid evolution of this element in these species; (2) the possibility that copies of the element are present only in nonsequenced heterochromatic regions of the genome; or (3) the stochastic loss of the element at some point in the evolution of these species.

Supplementary Material Table S1 summarizes the main information about the sequences found, such as the sequences used as queries and the nomenclature adopted in the present study. The sequences are available at http://www.ufsm.br/labdros/links/seqs.gypsy.

Sequence analysis

The *env* gene sequences used for the phylogenetic inferences included the following: (1) 27 *gypsy* sequences obtained from GenBank; (2) four sequences of the canonical *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* elements of *D. melanogaster* obtained from the Repbase Update (Jurka *et al.*, 2005); (3) 12 *gtwin* sequences (Ludwig & Loreto, 2007); (4) 10 *gypsy* sequences of the *flavopilosa* group species; and (5) 54 sequences of several retrotransposons obtained by genome search. In total, 107 sequences from 33 Drosophilidae (Fig. 1) species were used.

Two phylogenetic inference methods were used: Neighbor-Joining and Bayesian analysis, which revealed similar topologies. Figure 2 shows the tree obtained by the Bayesian analysis. According to the criterion of 20% nucleotide divergence applied by Herédia *et al.* (2004) to establish gypsy subfamilies, the phylogeny was divided into 17 main clades. These clades correspond to sequences homologous to the *gypsy4*, *gypsy3*, *gypsy6* and *gypsy2* retrotransposons. The remaining 13 groups have a monophyletic origin, of which seven (1, 2, 7, 9, 10, 11 and 12) correspond to the

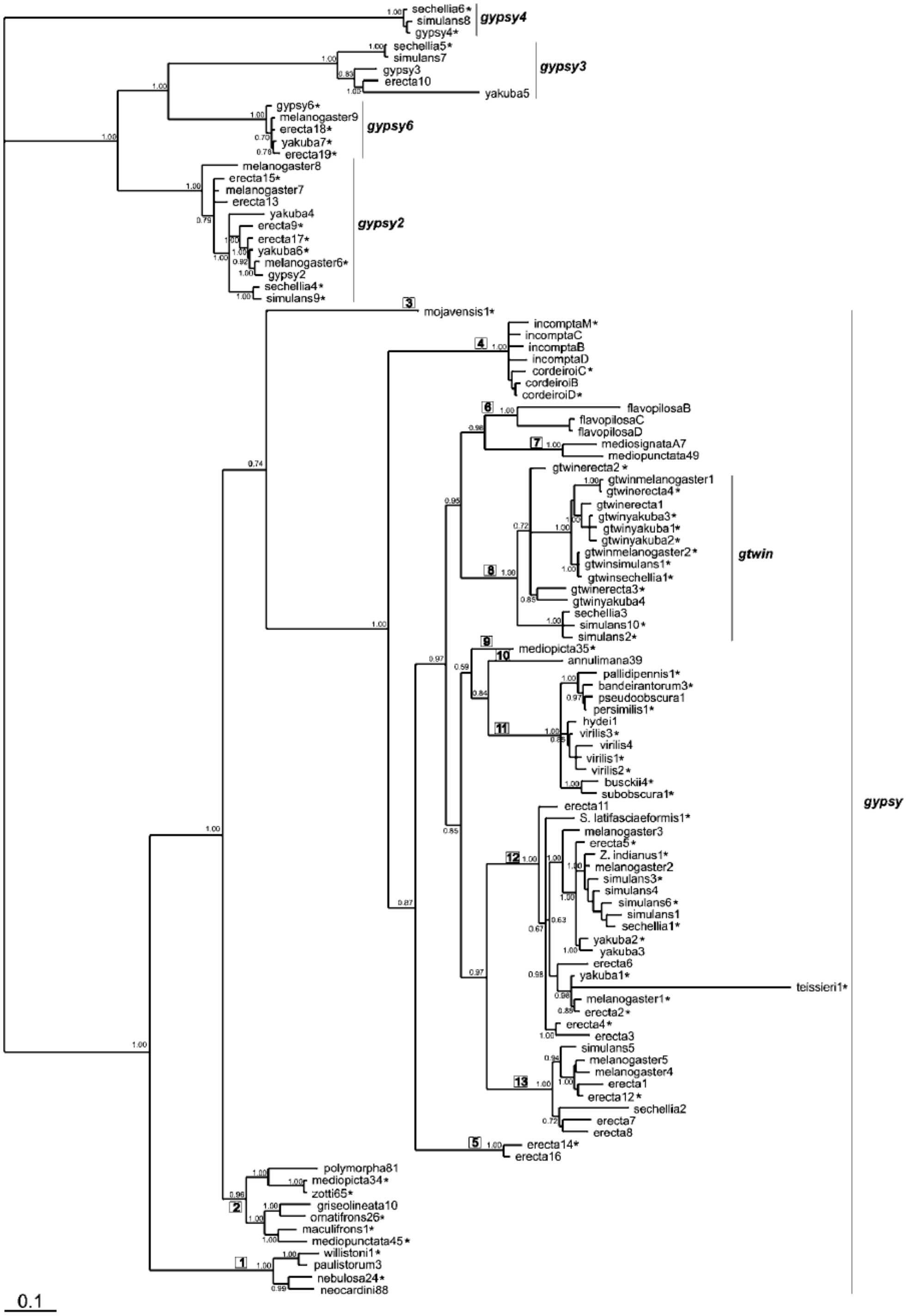


Figure 2. Phylogeny of *Errantivirus* using the Bayesian analysis with the GTR + G + I model for 107 sequences of approximately 500 bp of the *env* gene. Posterior probabilities for the major clades are shown next to the branches. The figure shows the divisions into different families of retrotransposons and *gypsy* subfamilies. Sequences marked with an asterisk (*) remained at the correct reading frame and did not have premature stop codons.

different *gypsy* subfamilies (G, F, B, D, E, A and C, respectively) as described in Herédia *et al.* (2004), although new sequences were added to groups 1, 11 and 12.

The *env* gene region analysed in this study potentially encodes roughly 106 amino acids. Fifty-nine sequences (55% of the total number of sequences analysed) lack a premature stop codon, are in the correct reading frame and exhibit the encoding potential for this *env* gene region.

The mean nucleotide and amino acid divergences within and among groups are shown in Table 1. The mean divergence among groups varied from 17.1% (group 9 and group 10) to 51.2% (group 6 and *gypsy3*) for nucleotide sequences and from 13.1% (group 9 and group 12) to 65.6% (group 2 and *gypsy4*) for amino acid sequences. When the different retrotransposon families were compared, *gypsy* element groups 1 to 13 presented nucleotide divergences of over 40% in relation to the other four families. For amino acids, divergences were over 50% in almost all comparisons. The comparison between the *gypsy4* family and the *gypsy2*, *gypsy3* and *gypsy6* families also shows considerable nucleotide and amino acid divergences (45.1–47.2% and 54.5–58.5%, respectively). Nevertheless, nucleotide and amino acid divergences among the *gypsy2*, *gypsy3* and *gypsy6* families were lower (29.7, 34.0 and 36.6% for nucleotides and 28.8, 36.1 and 38.8% for amino acids). These divergence values are comparable to, or even lower than, those found in some comparisons among groups 1 to 13.

Evolutionary analyses

In order to infer the occurrence of HT events, the following criteria were adopted: (1) incongruence in the retroelement's phylogeny, inferred in the present study, in comparison with the previously described host species evolutionary relationships; (2) divergence in the synonymous sites (dS) significantly lower in the elements when compared to nuclear genes in the same species. The α-*methyldopa* (*amd*) gene was used for dS comparison because it presents the highest number of sequences available for the studied species.

The comparison of dS values was chosen to infer HT because dS values offer a measurement of neutral evolution in the absence of a strong codon usage bias. Therefore, the same proportion of transposable elements (TEs) and host gene synonymous substitutions should be expected when two species are compared. If a TE has been transmitted from one species to another via a recent HT, the pairwise comparison between them should present a dS value for the TE significantly lower than the dS for the host genes, reflecting the lower divergence time of the TE sequences when compared to the species divergence time (Silva & Kidwell, 2000). Nevertheless, differences in the magnitude of dS among genes have been found to be negatively correlated to the intensity of natural selection on synonymous codon usage (Shields *et al.*, 1988; Sharp & Li, 1989). Thus,

Table 1. Nucleotide (below and to the left of the diagonal line) and amino acid (above and to the right of the diagonal line) divergence percentages found between and within LTR-retrotransposon groups

		gypsy																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	*gypsy2*	*gypsy3*	*gypsy4*	*gypsy6*
gypsy	**1**	9.6/17.9	36.6	42.9	47.1	42.1	–	–	41.8	41.8	–	42.1	42.0	42.3	51.4	57.3	59.9	49.1
	2	31.4	13.4/9.4	26.0	40.2	36.4	–	–	35.2	33.4	–	33.6	35.3	35.9	49.3	56.4	65.6	49.0
	3	37.0	28.3	nc/nc	39.5	38.8	–	–	39.6	38.4	–	38.1	36.4	37.8	49.7	56.3	63.6	49.8
	4	39.2	32.8	34.1	4.7/8.6	28.1	–	–	28.2	30.2	–	31.8	30.4	33.2	55.0	55.2	63.7	54.3
	5	37.6	32.5	35.7	28.5	3.3/nc	–	–	21.1	17.8	–	21.1	19.9	18.7	52.2	55.6	61.8	52.5
	6	37.3	35.3	37.5	31.2	30.0	13.4/–	–	–	–	–	–	–	–	–	–	–	–
	7	37.0	35.7	35.2	30.9	27.8	24.0	10.8/–	–	–	–	–	–	–	–	–	–	–
	8	35.8	32.1	33.7	28.1	26.4	25.7	25.2	10.6/9.8	16.4	–	20.7	19.2	20.7	52.1	55.4	63.2	54.1
	9	37.8	30.4	35.2	25.9	20.4	21.2	22.7	21.6	nc/nc	–	14.7	13.1	14.7	50.7	53.5	60.4	53.6
	10	37.3	33.7	35.9	28.1	23.7	27.2	26.2	22.9	17.1	nc/–	–	–	–	–	–	–	–
	11	36.1	31.6	33.6	28.3	26.8	27.5	25.8	24.1	20.1	21.1	6.9/4.6	18.4	20.2	51.4	52.6	60.9	52.5
	12	37.8	33.3	36.5	28.9	24.1	27.5	27.6	25.1	19.9	24.0	24.5	10.5/5.9	13.5	49.6	55.0	63.1	53.8
	13	39.1	34.2	35.4	30.6	24.4	27.4	27.3	25.3	20.7	23.5	24.7	22.0	9.3/nc	50.4	53.6	62.2	51.9
gypsy2		42.0	42.4	43.2	45.2	43.3	46.0	45.6	43.9	44.7	46.0	43.8	43.8	45.4	8.0/11.6	38.8	58.3	28.8
gypsy3		44.9	46.5	47.1	47.4	48.2	51.2	48.1	48.5	48.4	47.8	47.2	48.4	48.1	36.6	14.0/nc	54.5	36.1
gypsy4		47.9	48.7	48.3	48.1	47.8	48.5	50.4	47.2	47.7	47.9	47.8	48.3	49.2	47.2	46.6	1.2/0.6	58.5
gypsy6		43.3	42.7	41.4	45.6	45.1	48.9	47.2	46.9	47.0	47.3	47.8	47.5	47.5	29.7	34.0	45.1	1.5/2.7

nc, not calculated, only one sequence is available in the group; –, not calculated, at least one group does not have any sequences putatively encoding of *env* gene.

Table 2. Mean codon bias index (CBI) and number of effective codons (Nc) for *amd* gene and retrotransposon sequences. Nc varies between 21 (for maximum bias) and 61 (for minimum bias) and CBI varies between 0 (no bias) and 1 (maximum bias)

	CBI	Nc
amd	0.504	42.7
gypsy	0.359	54.9
gypsy2	0.343	52.5
gypsy3	0.315	57.9
gypsy4	0.279	58.2
gypsy6	0.398	58.8

we estimated the codon usage bias for retrotransposons and for the *amd* gene (Table 2).

It was observed that the *amd* gene is very useful for the comparisons between dS, as it does not present a very high codon usage bias. If this were the case, very low dS values would be observed, which would in turn lead to an underestimation of HT occurrence. Apart from this, the retrotransposons presented a lower degree of codon usage bias than the *amd* gene, which results in higher dS values and avoids the overestimation of HT occurrence. However, other phenomena may be held responsible for the distortions in the measurement of dS; eg, the sites required in splicing mechanisms or those involved in RNA secondary structure and other aspects linked to functionality of RNAs (Parmley *et al*., 2006; Xing & Lee, 2006).

In general, the comparison between the phylogeny obtained for retroelements and the phylogenetic relationships of the host species (Fig. 1) shows considerable incongruences that could suggest the occurrence of HT. These incongruences were tested through comparative dS analyses, using the Fisher's exact test. Figure 3 shows the main comparisons for each group.

The *gypsy4*, *gypsy3*, *gypsy6* and *gypsy2* families comprise only sequences of the *D. melanogaster* subgroup species. The clade topology of these elements is inconsistent with the host species phylogeny, which could be explained by ancestral polymorphism, stochastic loss or differences in evolutionary rates among sequences. However, the dS values comparisons (Fig. 3A,C,D) suggested that some of these incongruities might be the result of HT events.

The remaining sequences clustered as one large clade that corresponds to the *gypsy* family, divided into 13 subfamilies. The general analysis of these sequences reveals also vast incongruencies, as already proposed by Herédia *et al*. (2004), who suggested nine HT events.

The group 1 data analysis (Fig. 3E) indicates *gypsy* vertical transmission (VT) in the *D. willistoni* subgroup species, *D. willistoni* and *D. paulistorum* and two possible HT cases involving *D. nebulosa* and *D. neocardini*.

The topology of *gypsy* group 2 shows incongruities that are also detected by the dS comparisons (Fig. 3F). Although *D. medipicta* and *D. zotti* do not have nuclear gene sequences available for comparison, the low dS value found for *gypsy* (0.003) is difficult to reconcile with the hypothesis of VT. This group probably evolved by VT in the species *D. medipicta*, *D. ornatifrons* and *D. medipunctata*; however, three HT cases may have occurred: (1) from *D. ornatifrons* to *D. griseolineata* and (2) from *D. mediopunctata* to *D. maculifrons* and (3) from *D. mediopicta* to *D. zotti*.

The *gypsy* group 3 consists of only one *D. mojavensis* sequence, which belongs to the *repleta* group of the *virilis–repleta* radiation.

Group 4 is formed by *gypsy* sequences from *D. incompta* and *D. cordeiroi*, and the dS comparison (Fig. 3G) suggests that this element was vertically transmitted in these species.

The *gypsy* group 5 is composed by two *D. erecta* sequences that present a mean divergence of approximately 24% with the sequences of groups 12 and 13, which include other *D. erecta* sequences. This high divergence between sequences of the same species, as well as the position of group 5 in the phylogeny suggests that these sequences may have been acquired via HT from species that were not sampled in this analysis or that they are vestigial traces of an ancestral polymorphism that occurred long ago and was followed by stochastic loss in other related species.

The dS values comparison between *D. flavopilosa* (from group 6) and *D. mediopunctata* (from group 7) *gypsy* sequences (Fig. 3H) indicates that either (1) a very ancient HT event may have occurred between species of the *flavopilosa* and *tripunctata* groups, or (2) more likely, a recent HT event occurred from some species, which was not sampled in the present study, to the *D. flavopilosa* genome.

Group 8 is formed by sequences belonging to the *gtwin* family. The comparisons between dS values (Fig. 3I) lead to the same three possible HT events postulated recently by Ludwig & Loreto (2007). One of these events (between *D. melanogaster* and *D. erecta*) has already been corroborated by Kotnova *et al*. (2007).

The *gypsy* groups 9 and 10 correspond to only one sequence each, from *D. mediopicta* and *D. annulimana*, respectively. The *gypsy* dS value found for these two sequences was 0.531. Nuclear gene sequences for these two species are not available; however, by comparing the *amd* gene in *D. aracatacas* (a species of the *annulimana* group) and other species of the *quinaria–tripunctata* radiation, the dS values found were similar, varying between 0.568 and 0.660.

Data analysis of group 11 (Fig. 3J) indicates that this *gypsy* subfamily is possibly present in the *obscura* group species by means of VT. Nevertheless, the relationships among sequences of the remaining species reveal clear incongruencies with the host's phylogeny. Although *amd* sequences were not available for all species, the TE dS values found in some pairwise comparisons are much smaller than expected for species that belong to different

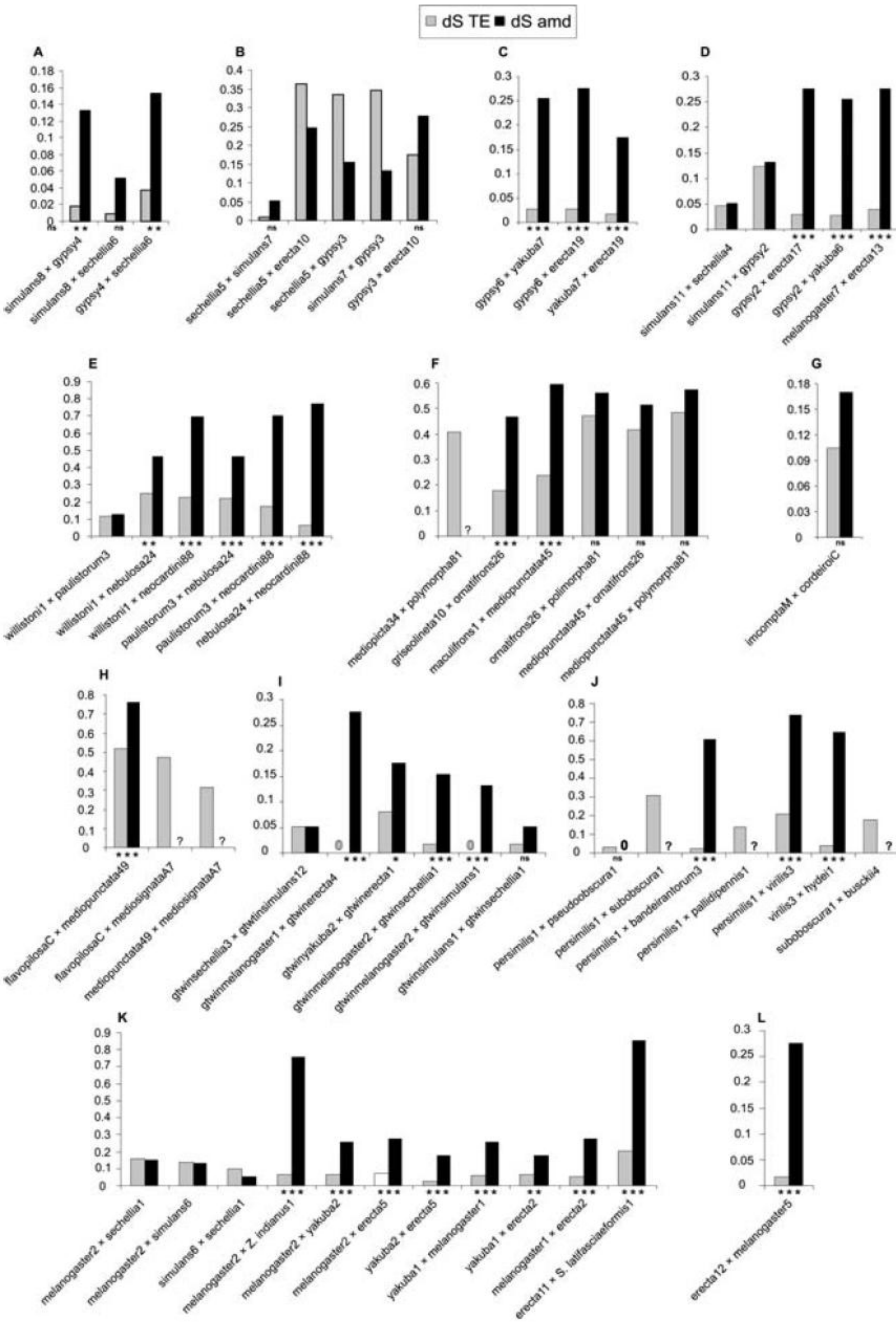


Figure 3. Comparative analysis of the dS values between retrotransposon and *amd* gene sequences. In the comparisons for which the dS value was zero, a zero (0) is shown in the figure, while in the comparisons for which it was not possible to calculate the dS, a question mark (?) is shown. The comparisons suggesting the occurrence of HT were tested by the Fisher's exact test. (*) – $P < 0.05$; (**) – $P < 0.01$; (***) – $P < 0.001$; (ns) – nonsignificant; A – comparisons for *gypsy4*; B – comparisons for *gypsy3;* C – comparisons for *gypsy6;* D – comparisons for *gypsy2*; E – comparisons for *gypsy* group 1; F – comparisons for *gypsy* group 2; G – comparisons for *gypsy* group 4; H – comparisons for *gypsy* groups 5 and 6; I – comparisons for *gypsy* group 8; J – comparisons for *gypsy* group 11; K – comparisons for *gypsy* group 12; L – comparisons for *gypsy* group 13.

Table 3. Summary of the probable horizontal transfer (HT) events identified, with the possible directions and times of occurrence

	Family/subfamily	Species 1	Direction	Species 2	Mya	Reference	Alteration
1	*gypsy/1*	paulistorum3	↔	neocardini88	5.437	a	A
2	*gypsy/1*	nebulosa24	↔	neocardini88	1.906	a	N
3	*gypsy/2*	ornatifrons2	→	griseolineata10	4.937	–	–
4	*gypsy/2*	mediopicta34	→	zotti65	0.094	a	N
5	*gypsy/2*	medipunctata45	→	maculifrons1	6.656	–	–
6	*gypsy* 6–7	mediosignataA7	↔	flavopilosaC	14.844	–	–
7	*gtwin-gypsy/8*	gtwinyakuba1	↔	gtwinerecta1	2.5	b	N
8	*gtwin-gypsy/8*	gtwinmelanogaster2	↔	gtwinsechellia1	0.531	b	N
9	*gtwin-gypsy/8*	gtwinmelanogaster1	↔	gtwinerecta4	0.0	b	N
10	*gypsy/11*	subobscura2	→	busckii4	5.531	a	N
11	*gypsy/11*	persimilis1	→	virilis3	6.375	a, c, d	A
12	*gypsy/11*	persimilis1	→	pallidipennis1	4.344	a*	A
13	*gypsy/11*	persimilis1	→	bandeirantorum3	0.812	a*	A
14	*gypsy/11*	virilis3	→	hydei1	1.156	a, d	A
15	*gypsy/12*	erecta11	→	S.latifasciaeformis1	6.344	a	A
16	*gypsy/12*	melanogaster2	→	Z.indianus1	2.0	a	A
17	*gypsy/12*	melanogaster2	→	yakuba2	2.0	–	–
18	*gypsy/12*	melanogaster1	↔	yakuba1	1.937	–	–
19	*gypsy/12*	melanogaster1	↔	erecta2	1.687	–	–
20	*gypsy/12*	yakuba2	→	erecta5	0.875	–	–
21	*gypsy/13*	melanogaster5	↔	erecta12	0.531	–	–
22	*gypsy2*	melanogaster7	↔	erecta15	1.219	–	–
23	*gypsy2*	gypsy2	→	erecta17	0.906	–	–
24	*gypsy2*	gypsy2	→	yakub6	0.844	–	–
25	*gypsy4*	simulans8	↔	gypsy4	0.563	–	–
26	*gypsy6*	gypsy6	↔	erecta19	0.843	–	–
27	*gypsy6*	erecta19	↔	yakuba7	0.562	–	–

a – Cases reported by Herédia *et al.* (2004);
a* – Case reported by Herédia *et al.* (2004) as just one HT event from *Drosophila bandeirantorum* to *D. pallidipennis*;
b – Cases reported by Ludwig & Loreto (2007);
c – Case reported by Alberola & de Frutos (1996);
d – Cases reported by Vázquez-Manrique *et al.* (2000), with *D. afinnis* as donor species;
A, HT described events with alteration of donor–receptor species;
N, HT described events cases without alteration of donor–receptor species.

subgenera, separated by roughly 60 Myr (Tamura *et al.*, 2004). Therefore, five possible HT events were suggested in this *gypsy* group.

Group 12 shows several topological inconsistencies and a complex relationship among sequences of the *melanogaster* subgroup species, with several HT events (Fig. 3K). Moreover, *gypsy* HT may explain the presence of *Zaprionus indianus* and *Scaptodrosophila latifascieformis* sequences in this group.

Similarly to group 12, group 13 also has sequences of the *D. melanogaster* subgroup species. An estimate of the dS value was made only for the melanogaster5 and erecta12 sequences (Fig. 3L), as the other sequences had many indels. This comparison suggests an HT event between *D. melanogaster* and *D. erecta*. The unexpected topology of the branches of this group may also be explained by ancestral polymorphism that was followed by sampling and/or by stochastic loss.

The data presented in this study suggest the occurrence of 27 possible HT events, of which 15 have not been reported before. In some previously described cases, donor and recipient species changed with the inclusion of the new sequences. For most cases, the estimated time of transfer occurrence was under 1 Myr. Table 3 summarizes the probable HT events detected, with the possible directions and times.

Discussion

Errantivirus diversity in the genomes

No general consensus exists concerning the criteria for TE classification, especially regarding the definition of families and subfamilies. Several authors have used different amino acid and/or nucleotide divergence percentages to classify distinct TEs (Lohe *et al.*, 1995; Robertson, 1995; Clark & Kidwell, 1997; Capy *et al.*, 1998; Bowen & McDonald, 2001; Herédia *et al.*, 2004). Thus, we chose to keep the initial status of each sequence group (different retrotransposon families and different *gypsy* subfamilies), while the new groups, because of their position in the phylogeny, were considered *gypsy* subfamilies.

Several studies have shown the coexistence of different subfamilies or variants of elements in the genomes [*mariner* (Robertson & MacLeod, 1993), *gypsy* (Terzian *et al.*, 2000; Herédia *et al.*, 2004) and *P* element (Clark & Kidwell, 1997), among others]. This arrangement was also observed in our study, especially for the species of the *melanogaster* subgroup, which have representative sequences of different families, subfamilies and variants of the same subfamily coexisting in the same genome. Two possible situations may explain these observations: (1) ancestral polymorphism or (2) successive invasions by these errantiviruses into the genomes.

Phylogeny based on the env *gene places the* gtwin *retroelement within the* gypsy *phylogeny*

The *gtwin* retroelement was discovered by *in silico* analyses of the *D. melanogaster* genome (Bowen & McDonald, 2001). The data currently available show the restricted distribution of this element in the *melanogaster* subgroup species (Kotnova *et al*., 2007; Ludwig & Loreto, 2007). The *gag*, *pol* and *env* genes of *gtwin* show 53, 76, and 77% identity, respectively, in amino acid sequences with the *gypsy* retrotransposon; however, the LTR sequences of these two elements show little similarity (Kotnova *et al*., 2005). According to Kotnova *et al*. (2005), the *gtwin* may have just recently acquired its *env* gene via recombination with *gypsy*, as these elements are largely similar in this gene, which is not expected for the comparisons of different retroelements. However, although our data show sequences homologous to the *gtwin env* gene within the *gypsy* clade, these are related to *gypsy* sequences of the subgenus *Drosophila* species, rather than to the *D. melanogaster gypsy* subfamilies. Thus, it is possible to suggest that the *gtwin* retroelement emerged during the evolution of the subgenus *Drosophila*, undergoing a recombination event with sequences of the *gypsy* retroelement of these species. The presence of this element in the *D. melanogaster* subgroup species may be explained by HT. More comprehensive studies on the distribution of the *gtwin* retroelement could shed new light on its evolutionary history.

The complex evolution of errantiviruses with multiple HT events

In this investigation on the evolutionary relationships of several *Errantivirus* sequences in Drosophilidae species, various incongruences between TE and host species phylogenies were observed. Most of these inconsistencies may be explained by HT. Nevertheless, it is difficult to infer the direction of the majority of the movements postulated. Other explanations may still assist in clarifying the complex branch topologies of the retroelements' phylogeny, such as different evolution rates among species, ancestral polymorphism, different distributions of polymorphic copies during speciation and stochastic loss. Acting concomitantly, these different evolutionary factors may have shaped the evolution of errantiviruses in the genus *Drosophila*.

Nine HT events had previously been described for the *gypsy* retroelement, which is broadly distributed (Alberola & de Frutos, 1996; Vázquez-Manrique *et al*., 2000; Herédia *et al*., 2004). With the inclusion of new *gypsy* sequences and with the availability of *amd* gene sequences for the majority of the species analysed, we were able to postulate nine additional HT events for this element.

We have found that *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6*, as well as *gtwin* (Kotnova *et al*., 2007; Ludwig & Loreto, 2007), seem to be restricted to the *melanogaster* subgroup. Nonetheless, it is important to take into consideration the fact that a large number of this group's species have genome sequences available. Thus, a more far-reaching analysis could change this scenario. For *gypsy2*, *gypsy4* and *gypsy6* at least one possible HT event was postulated for each of them. Based on the phylogenetic results, it is possible to suggest that these elements are prone to invade one or more genomes of this subgroup species through multiple HT waves of unknown origins. This is in accordance with some previous indications that the majority of TEs in *D. melanogaster* are of recent origin, possibly as a result of the arrival of TE families via HT (Bowen & McDonald, 2001; Sanchez-Gracia *et al*., 2005). For *gypsy3*, although there was a topological inconsistency, HT was not corroborated by the dS comparisons.

Horizontal transfer, which initially was considered a rare phenomenon, has proven its significant relevance, at least in TE studies. Even though several TE HT events have been documented, we do not know the details surrounding these processes. Some potential vectors for transmission are parasitic wasps, mites and endosymbiotic bacteria (Silva *et al*., 2004). The retrotransposons could simply be copackaged within the viral capsid of RNA viruses, which can infect different host species (Malik *et al*., 2000). In addition, HT requires not only that the distributions of donor and recipient overlap in a geographic sense, but it is probably facilitated by ecological and temporal overlap as well (Silva *et al*., 2004). These requirements are satisfied for the species involved in possible HTs of this work, corroborating our results.

We observed a considerable number of the potential HT events between species of the *melanogaster* subgroup. Other studies also showed a high number of HT events among these species, both for retroelements (Sanchez-Gracia *et al*., 2005; Ludwig & Loreto, 2007) and DNA transposons (Sanchez-Gracia *et al*., 2005). This fact poses the question of the possible role of introgression played in the acquisition of new sequences. Introgression phenomena were proposed by Silva & Kidwell (2000) as a potential mechanism for the *P* element spread among the *D. willistoni* subgroup species. This process may have occurred in the *melanogaster* subgroup species, mainly among *D. melanogaster*, *D. simulans* and *D. sechellia*, which are able to generate fertile hybrids, depending on the combinations between parental lineages (Lachaise *et al*., 1986; Davis *et al*., 1996). Nevertheless, for the other species of this subgroup that present a greater divergence time, this explanation is less probable.

Errantiviruses, however, seem to be rather prone to HT and several cases between distantly related species have been reported (Alberola & de Frutos, 1996; Vázquez-Manrique *et al.*, 2000; Herédia *et al.*, 2004). The acquisition of an *env* gene releases these elements from relying on another vector for jumping into different hosts, because then they could exhibit an autonomous infectious capacity, increasing the probability of cross-species transfer (Malik *et al.*, 2000; Mejlumian *et al.*, 2002; Herédia *et al.*, 2004). The high retrotransposon HT rate found in our study agrees with this hypothesis. In most HT events postulated, the sequences involved have retained the capacity to encode the *env* gene region. Thus, in spite of the fact that, up until now, only the infectious potential of *gypsy* in *D. melanogaster* had been investigated and demonstrated, *gypsy* sequences of other species and other *Errantivirus* may also be infectious or may have been in the recent past. The large number of HT events among the *melanogaster* subgroup species may be explained, in this context, by a more frequent transmission rate amongst closely related species, which is made clear by the study of vertebrate retroviruses (Martin *et al.*, 1999; Gifford, 2006).

Finally, considering that TEs have direct implications in the genome evolution, in order to understand the host-TE coevolution, it is also important to know, besides the extension of the HT events, the mechanisms of these DNA transfers and the effects of this process in the host genomes.

Evolution of Errantivirus

Malik *et al.* (2000) showed that the *env* gene has been acquired independently on several instances in the evolutionary history of LTR retrotransposons and suggested that these acquisitions may increase the probability of HT events. One of these acquisitions occurred in the genus *Errantivirus*, which forms a monophyletic group of insect retrotransposons (Malik *et al.*, 2000; Terzian *et al.*, 2001). They evolved from a single recombination event, through which the progenitor incorporated one *env*-like open reading frame from a baculovirus, which are pathogenic insect viruses (Malik *et al.*, 2000; Misseri *et al.*, 2004).

Thus, it is possible to conjecture that the ancestral sequence of *Errantivirus* exhibited infectious capacity and that mutation and recombination events throughout the evolutionary history of host species generated the different *Errantivirus* species. The dissemination of *Errantivirus* may have happened via VT of copies that were incorporated in the germline and via multiple HT events across species. During the evolution of these elements, several of them may have lost their infectious capacity and started to behave just like endogenous retroviruses. Nevertheless, the results of the present study suggest that the evolution of these sequences was marked by the conservation of the *env* gene in several *Errantivirus* species. The preservation of this gene leads to the maintenance of their infectious capacity, indicating that this may be decisive to the survival and expansion of these retroviruses in the genomes.

Experimental procedures

Species and origin of sequences

A total of 31 *Drosophila* species were used in this study, together with *Zaprionus indianus* and *Scaptodrosophila latifasciaeformis* (Fig. 1). For some of these species, sequences deposited in GENBANK (Benson *et al.*, 2007) for the *env* gene region of *gypsy* (position 6026–6511 of the *gypsy* element in *D. melanogaster*) and the *gtwin* sequences obtained at http://www.ufsm.br/labdros/links/seqs.fasta (Ludwig & Loreto, 2007) were used. New sequences were obtained for the *flavopilosa* group species by cloning and sequencing and for other *Drosophila* species by searching the available genomes sequences (see below).

PCR, cloning and sequencing

The species of the *flavopilosa* group were collected from *Cestrum parqui* and *C. calycinum* flowers. Specimen identification was conducted as in Brncic (1978). DNA from *D. flavopilosa*, *D. incompta* and *D. cordeiroi* was prepared from adult males according to Sassi *et al.* (2005). Degenerate primers described in Herédia *et al.* (2004) were used to amplify the approximately 485-bp fragment of the *env* gene. The components of the 50 µl reaction mixture were 100 ng DNA, 1 U Taq polymerase, 5 µl 10× reaction buffer supplied by the manufacturer (Invitrogen, Carlsbad, CA), 200 µM of each nucleotide, 20 pmol of each primer and 1.5 mM $MgCl_2$. Amplification parameters were 96 °C for 2 min, 35 cycles at 96 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min, followed by an extension cycle at 72 °C for 7 min. PCR products were cloned into the PCR2.1 cloning vector using the TA Cloning® kit (Invitrogen, Carlsbad, CA, USA). DNA sequencing was carried out directly from the purified plasmids in the MegaBace 500 automatic sequencer. The dideoxy chain-termination reaction was implemented with the use of the DYEnamicET kit (GE Healthcare, Little Chalfont, Bucks, U.K.). Both DNA chains were sequenced. Consensus sequences for each clone were obtained by the electropherogram analyses using the GAP 4 software of the Staden Package (Staden, 1996).

Searches in genomes

Searches were conducted for *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. virilis*, *D. mojavensis* and *D. grimshawi* using the BLAT tool (http://genome.ucsc.edu/cgi-bin/hgBlat; Kent, 2002) and for *D. willistoni* using the BLAST tool (http://flybase.bio.indiana.edu/blast/; Grumbling & Strelets, 2006). The sequences used as query were one sequence of each *gypsy* subfamily established by Herédia *et al.* (2004), the sequences of the *gypsy2*, *gypsy3*, *gypsy4* and *gypsy6* families of *D. melanogaster* (Repbase Update, Jurka *et al.*, 2005) and the sequences of the *flavopilosa* group species. Some of the sequences found in these searches were also used as query. For each species, all sequences found with a score > 230 were pairwise aligned using GENEDOC 2.6.003 software (Nicholas *et al.*, 1997). In order to obtain different *gypsy* variants for the same species without using a very high number of sequences, we maintained only sequences presenting divergence over 2% as compared to the other sequences of the same species. The sequences obtained had their identity ascertained by CENSOR software (Kohany *et al.*, 2006).

Phylogenetic analysis, dS estimates and preferential codon usage

Nucleotide sequences were aligned using ClustalX software (Thompson *et al*., 1997). The GTR + G + I evolutionary model was appointed by MrModelTest 2.2 (Nylander, 2004), according to the Akaike information criterion (Akaike, 1974). Two phylogenetic analysis methods were used: (1) Neighbor-Joining (Saitou & Nei, 1987), run in Mega 3.1 software (Kumar *et al*., 2004) and using the Tamura–Nei model with gamma distribution values indicated by MrModelTest 2.2 (Nylander, 2004) and a 1000-replication bootstrap and (2) Bayesian analysis using MrBayes 3.1.2 software (Ronquist & Huelsenbeck, 2003), with evaluation of at least 270 000 generations and a burnin region of 675 trees using the model appointed by MrModelTest 2.2 (Nylander, 2004). The posterior probability for each clade was obtained by the 50% majority rule of the tree consensus (Hall, 2001).

The amino acid sequences were deduced using GeneDoc 2.6.001 software (Nicholas *et al*., 1997). The sequences that remained at the correct reading frame and that did not have premature stop codons were used to calculate the divergence between amino acid sequences using the P-distance. The P-distance was also used to calculate the divergence between the nucleotide sequences.

A codon alignment for each group was used to estimate the dS values and the number of synonymous sites (S) using the Nei & Gojobori (1986) method assisted by the Mega 3.1 software (Kumar *et al*., 2004). In order to allow the calculation of some sequences, gaps were introduced to conserve the reading frame while stop codons were considered as absent information.

The effective number of codons (Nc; Wright, 1990) and the codon bias index (CBI; Morton, 1993) were calculated by DnaSP 4.0 software (Rozas *et al*., 2003).

Inference of horizontal transmission

In order to infer HT, we considered (1) the incongruities existing between phylogenetic relationships of retroelements and the evolutionary relationships of the host species and (2) the lower dS value between sequences of the TE compared to the *amd* nuclear gene. Accession numbers of the *amd* sequences used in this study are available in Supplementary Material Table S2. Estimates of dS, CBI and Nc were also obtained for the *amd* gene as described above.

Thus, for each HT event possible, we resorted to Fisher's test, assisted by DnaSP 4.0 software, to verify whether the number of similarities and differences observed for TEs in synonymous sites were significantly lower than those observed for *amd* gene.

Only one HT case was inferred when two or more related species are involved in the same event. Thus, in each group of phylogenetically related species, the one most similar to the distant species (receptor) was considered the donor species.

The estimated time that HT events occurred was conducted according to the formula $T = k/2r$ (Graur & Li, 2000), in which T is the divergence time between species, k is the divergence between TE sequences (dS) in synonymous sites and r is the evolutionary rate. We used a synonymous substitution rate of 0.016 substitutions per site per million years, calculated for *Drosophila* genes with low codon usage bias (Sharp & Li, 1989).

Acknowledgements

This work was supported by grants from FAPERGS (0411965) and CNPq. We thank CNPq for fellowships. We thank Agencourt Bioscience Corporation, the Broad Institute, J. Craig Venter Institute and Washington University School of Medicine Genome Sequencing Center for prepublication access to the genome sequence data. We are grateful to M.Sc. Newton de Medeiros Vidal, M.Sc. Lizandra Jaqueline Robe and B.Sc. Nina Roth Mota for the valuable help and critical comments which contributed significantly to the improvement of this paper.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transaction on Automatic Control* **19**: 716–723.

Alberola, T.M. and de Frutos, R. (1996) Molecular structure of a *gypsy* element of *Drosophila subobscura* (*gypsyDs*) constituting a degenerate form of insect retroviruses. *Nucleic Acids Res* **24**: 914–923.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2007) GenBank. *Nucleic Acids Res* **35**: D21–25.

Boeke, J.D., Eickbush, T.H., Sandmeyer, S.B. and Voytas, D.F. (1999) Metaviridae. In *Virus Taxonomy: ICTV VII Report* (Murphy, F.A., ed.), pp. 123–135. Springer-Verlag, NY.

Bowen, N.J. and McDonald, J.F. (2001) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res* **11**: 1527–1540.

Brncic, D. (1978) A note on the *flavopilosa* group of species of *Drosophila* in Rio Grande do Sul, Brazil (Diptera: Drosophilidae) with the description of two new species. *Rev Bras Biol* **38**: 647–651.

Capy, P., Bazin, C., Higuet, D. and Langin, T. (1998) *Dynamics and Evolution of Transposable Elements*. Landes Bioscience, Austin, TX.

Clark, J.B. and Kidwell, M.G. (1997) A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proc Natl Acad Sci USA* **94**: 11428–11433.

Davis, A.W., Roote, J., Morley, T., Sawamura, K., Herrmann, S. and Ashburner, M. (1996) Rescue of hybrid sterility in crosses between *D. melanogaster* and *D. simulans*. *Nature* **380**: 157–159.

Gifford, R.J. (2006) Evolution at the host–retrovirus interface. *BioEssays* **28**: 1153–1156.

Graur, D. and Li, W.-H. (2000) *Fundamentals of Molecular Evolution*, 2nd edn. Sinauer Associates, Sunderland, MA.

Grumbling, G. and Strelets, V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res* **34**: D484–488.

Hall, B.G. (2001) *Phylogenetic Trees Made Easy: a How to Manual for Molecular Biologists*, 2nd edn. Sinauer Associates, Sunderland, MA.

Herédia, F., Loreto, E.L.S. and Valente, V.L. (2004) Complex evolution of *gypsy* in drosophilid species. *Mol Biol Evol* **21**: 1831–1842.

Herédia, F., Loreto, E.L.S. and Valente, V.L. (2007) Distribution and conservation of the transposable element *gypsy* in drosophilid species. *Genet Mol Biol* **30**: 133–138.

Jurka, J. (2000) Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**: 418–420.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogent Genome Res* **110**: 462–467.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S. *et al*. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* **3**: 0084.1–0084.20.

Kapitonov, V.V. and Jurka, J. (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci USA* **100**: 6569–6574.

Kent, W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Prud'homme, N. and Bucheton, A. (1994) Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **91**: 1285–1289.

Kohany, O., Gentles, A.J., Hankus, L. and Jurka, J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**: 474.

Kotnova, A.P., Feoktistova, M.A., Glukhov, I.A., Salenko, V.B., Lyobomirskava, N.V., Kimb, A.I. and Ilyina, Y.V. (2005) Retrotranspason *Gtwin* specific for the *Drosophila melanogaster* subgroup. *Dokl Biochem Biophys* **409**: 233–235.

Kotnova, A.P., Glukhov, I.A., Karpova, N.N., Salenko, V.B., Lyubomirskaya, N.V. and Ilyin, Y.V. (2007) Evidence for recent horizontal transfer of *gypsy*-homologous LTR-retrotransposon *gtwin* into *Drosophila erecta* followed by its amplification with multiple aberrations. *Gene* **396**: 39–45.

Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**: 150–163.

Lachaise, D., David, J.R., Lemeunier, F., Tsacas, L. and Ashburner, M. (1986) The reproductive relationships of *Drosophila sechellia* with *D. mauritiana*, *D. simulans* and *D. melanogaster* from the Afrotropical region. *Evolution* **40**: 262–271.

Lewis, R.L., Beckenbach, A.T. and Mooers, A.Ø. (2005) The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on *COI* and *COII* sequences and a Bayesian estimate of phylogeny. *Mol Phylogenet Evol* **37**: 15–24.

Lohe, A.R., Moriyama, E.M., Lidholm, D.A. and Hartl, D.L. (1995) Horizontal transmission, vertical inactivation, and stochastic loss of mariner like transposable elements. *Mol Biol Evol* **12**: 62–72.

Ludwig, A. and Loreto, E.L.S. (2007) Evolutionary pattern of the *gtwin* retrotransposon in the *Drosophila melanogaster* subgroup. *Genetica* **130**: 161–168.

Malik, H.S., Henikoff, S. and Eickbush, T.H. (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**: 1307–1318.

Martin, J., Herniou, E., Cook, J., O'Neill, R.W. and Tristem, M. (1999) Interclass Transmission and phyletic host tracking in murine leukemia virus-related retroviruses. *J Virol* **73**: 2442–2449.

Mejlumian, L., Pelisson, A., Bucheton, A. and Terzian, C. (2002) Comparative and functional studies of *Drosophila* species invasion by the *gypsy* endogenous retrovirus. *Genetics* **160**: 201–209.

Misseri, Y., Cerutti, M., Devauchelle, G., Bucheton, A. and Terzian, C. (2004) Analysis of the *Drosophila gypsy* endogenous retrovirus envelope glycoprotein. *J Gen Virol* **85**: 3325–3331.

Morton, B.R. (1993) Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on *tRNA* availability, *J Mol Evol* **37**: 273–280.

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

Nicholas, K.B. and Nicholas, H.B.J. (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author. http://www.psc.edu/biomed/genedoc

Nylander, J.A.A. (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University, Uppsala.

O'Grady, P.M. and Kidwell, M.G. (2002) Phylogeny of the subgenus *Sophophora* (Diptera: Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. *Mol Phylogenet Evol* **22**: 442–453.

Parmley, J.L., Chamary, J.V. and Hurst, L.D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **2**: 301–309.

Robe, J.L., Valente, V.L.S., Budnik, M. and Loreto, E.L.S. (2005) Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on Neotropical species and groups: a nuclear vs. mitochondrial gene approach. *Mol Phylogenet Evol* **36**: 623–640.

Robertson, H.M. (1995) The *Tc1-mariner* superfamily of transposons in animals. *J Insect Physiol* **41**: 99–105.

Robertson, H.M. and MacLeod, E.G. (1993) Five major subfamilies of mariner transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. *Insect Mol Biol* **2**: 125–139.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–425.

Sanchez-Gracia, A., Maside, X. and Charlesworth, B. (2005) High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet* **21**: 200–203.

Sassi, A.K., Herédia, F., Loreto, E.L.S., Valente, V.L.S. and Rohde, C. (2005) Transposable elements P and *gypsy* in natural populations of *Drosophila willistoni*. *Genet Mol Biol* **28**: 734–739.

Sharp, P.M. and Li, W.H. (1989) On the rate of DNA sequence evolution in *Drosophila*. *J Mol Evol* **28**: 398–402.

Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) 'Silent' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5**: 704–716.

Silva, J.C. and Kidwell, M.G. (2000) Horizontal transfer and selection in the evolution of P elements. *Mol Biol Evol* **17**: 1542–1557.

Silva, J.C., Loreto, E.L.S. and Clark, J.B. (2004) Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* **6**: 57–71.

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D. and Corces, V.G. (1994) An env-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. *Genes Dev* **8**: 2046–2057.

Stacey, S.N., Lansman, R.A., Brock, H.W. and Grigliatti, T.A. (1986) Distribution and conservation of mobile elements in the genus *Drosophila*. *Mol Biol Evol* **3**: 522–534.

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

Tamura, K., Subramanian, S. and Kumar, S. (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol* **21**: 36–44.

Terzian, C., Ferraz, C., Demaille, J. and Bucheton, A. (2000) Evolution of the *Gypsy* endogenous retrovirus in the *Drosophila melanogaster* subgroup. *Mol Biol Evol* **17**: 908–914.

Terzian, C., Pelisson, A. and Bucheton, A. (2001) Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol* **1**: 3.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.

Vázquez-Manrique, R.P., Hernandez, M., Martinez-Sebastian, M.J. and de Frutos, R. (2000) Evolution of *gypsy* endogenous retrovirus in the *Drosophila obscura* species group. *Mol Biol Evol* **17**: 1185–1193.

Wheeler, M.R., Takada, H. and Brncic, D. (1962) The *flavopilosa* species group of *Drosophila*. *Univ Texas Publ* **6205**: 395–413.

Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene* **87**: 23–29.

Xing, Y. and Lee, C. (2006) Can RNA selection pressure distort the measurement of Ka/Ks? *Gene* **370**: 1–5.

Supplementary material

The following supplementary material is available for this article:

Table S1 List of retroelements sequences obtained from GenBank and Repbase, with the respective accession number and retroelements sequences obtained by searches in the genomes, with the queries sequences and chromosomal localization

Table S2 Accession numbers for the *amd* sequences

This material is available as part of the online article from http://www.blackwell-synergy.com/doi/full/10.1111/j.1365-2583.2007.00787.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Capítulo 3

Evolution of the endogenous retrovirus *nik* in *Drosophila*

Adriana Ludwig[1], Vera L. S. Valente[1,2], Elgion L. S. Loreto[1,3]

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil.

[2] Departamento de Genética, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil.

[5] Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), CEP 97105-900, Santa Maria, Rio Grande do Sul, Brazil; phone 55-55-32208912; email – elgion.loreto@pq.cnpq.br.

Manuscrito em preparação para ser enviado à revista Genetica.

ABSTRACT

LTR-retrotransposons are the most abundant transposable elements in *Drosophila.* The LTR-retrotransposon *nik*, also known as *gypsy5*, was discovered by the analysis of the *D. melanogaster* genomic sequence it is evolutionarily closer to *gypsy-like* endogenous retroviruses. Aiming to understand the evolutionary history of *nik* we conducted *in silico* searches of this element in the twelve available *Drosophila* genomes. Additionally, searches were performed by PCR in a large number of species. The presence of divergent and degenerated copies suggests that *nik* is ancient in *Drosophila* genomes and was probably lost in major of species. Diversification of *nik* probably have occurred prior to the *melanogaster* subgroup speciation and different evolutionary process, as independent assortment of copies, stochastic lost, introgression, and horizontal transfer, help to explain the current distribution of *nik*. Moreover, we suggest that despite to have a general retrotransposon silencing mechanism in *Drosophila,* which probably controls *nik* by piRNAs, some components in this process can be different among species.

Keywords: *gypsy*, endogenous retroviruses, *env* gene, *flamenco*

INTRODUCTION

LTR-retrotransposons are among the most abundant constituents of eukaryotic genomes and are evolutionarily closely related to vertebrate retroviruses. Their integrated proviral form consists of two long terminal repeats (LTRs) that flank the internal coding region, which includes genes encoding both structural and enzymatic proteins (Havecker et al. 2004).

Most LTR-retrotransposons have only two genes: *gag*, which encodes structural proteins that form the virus-like particle and *pol*, which contains the information for the Reverse Transcriptase (RT), Protease (PR) and Integrase (IN) enzymes. These genes are believed to be necessary and sufficient for transposition and intracellular life cycle of LTR-retrotransposons. Retroviruses, however, have an extracellular life that is mediated by the presence of an envelope (*env*) gene that encodes the surface and transmembrane components of the viral Env protein (Coffin et al. 1997).

Based on the organization of domains and the reverse transcriptase phylogeny, LTR-retroelements can be divided into four major groups: three retrotransposon groups, *Ty1-copia*, *Ty3-gypsy*, *Bel-Pao* and the vertebrate retroviruses (Eickbush e Jamburuthugoda 2008). However, the presence of an *env* gene is not restricted to the vertebrate retroviruses. Several instances of *env-like* gene acquisitions have taken place in the LTR-retroelements evolutionary history (Malik et al. 2000).

A particular insect retrotransposons evolutionary lineage, from *Ty3-gypsy* group, contains an *env* gene acquired from insect baculoviruses (double-stranded DNA viruses) (Malik et al. 2000). This group is highly diverse with several families described in different species (Bowen and McDonald 2001; Kaminker et al. 2002; Kapitonov and Jurka, 2003). The *gypsy* element from *D. melanogaster* is the only one that has shown infectious properties (Kim et al. 1994; Song et al. 1994). However, several other families also display the structural requirements for infectiousness, such as the capability to encode an Env protein (Leblanc et al. 2000; Llorens et al. 2008; Ludwig et al. 2008; Mejlumian et al. 2002). These retrotransposons are known as insect endogenous retroviruses, or also *gypsy-like* elements.

Several works have shown that *gypsy-like* elements are frequently involved in horizontal transfer (HT) events (Herédia et al. 2004; Ludwig and Loreto 2007; Ludwig et al. 2008; Llorens et al. 2008; Bartolomé et al. 2009; Terzian et al. 2000; Vidal et al. 2009). We have suggested that HT may be a decisive process in the evolution, expansion and maintenance of endogenous retrovirus in *Drosophila* (Ludwig et al. 2008; Vidal et al. 2009).

The spread of some *Drosophila* endogenous retroviruses such as *gypsy, ZAM* and *idefix* is controlled by *flamenco* locus that maps to the pericentromeric heterochromatin on the X chromosome (Prud'homme et al. 1995; Desset et al. 2003). The *flamenco* locus consists of a large number of truncated or defective retrotransposons and is a source of piRNAs that are involved in retrotransposon silence in somatic cells of ovary (Brennecke et al. 2007; Malone et al. 2009).

The insect endogenous retroviruses *nik*, also named *gypsy5* was discovered by the *D. melanogaster* genome analysis (Bowen and McDonald 2001). We investigate the *nik* evolution using the 12 *Drosophila* genomes and PCR approach and we also discuss the evolution of this endogenous retrovirus in the context of *flamenco* locus.

MATERIAL AND METHODS

Genome searches

We used the BLAST tool available on the Flybase (http://flybase.bio.indiana.edu/blast/; Grumbling and Strelets 2006) to perform searches in the 12 *Drosophila* genomes. We first conducted BLASTn using the complete canonical *nik* (*gypsy5*) (Kapitonov and Jurka 2000) as query. Sequences with expected value < 10^{-4} and covering at least 50% of at least one gene region (*gag*, *pol* and *env*) were considered and used in the phylogenetic analyses. Censor software (Kohany et al. 2006) was used to help identifing the boundaries of *nik* copies. Additionally, tBLASTx was used to identify more divergent *nik* copies that could be present in more divergent species.

The *flamenco* genomic regions identified for *D. melanogaster* (chrX: 21,492,100 – 21,687,300), *D. yakuba* (chrX: 20,617,000 – 20,724,800) and *D. erecta* (scaff.4690: 17,958,500 – 18,037,650) (Malone et al. 2009) were screened by Censor software (Kohany et al. 2006) to identify the *nik* regions that are present in this locus.

nik gene organization

Open reading frames (ORFs) were identified using ORF finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The gag-pol expression strategy was inferred by comparing the *gag-pol* overlapping region of *nik* with related retrotransposons that have the putative frameshifting site already identified (Gao et al. 2003).

Cloning and sequencing

To further define the distribution of *nik* and to infer phylogenetic relationships, a survey by PCR amplification was performed in 65 Drosophilidae species (table S1 from supplementary material). DNA was extracted from 30 fresh adult flies according to Sassi et al. (2005). Amplified samples were purified using GFX Purification Kit (GE Healthcare) and cloned with TOPO-TA cloning vector (Invitrogen). The clones were sequenced from the purified plasmids using the universal primers, M13 forward and M13 reverse, in Megabace 500 sequencer. The dideoxy chain-termination reaction was implemented using the DYEnamicET kit (GE Healthcare). Both DNA strands were sequenced at least twice or until to obtain very reliable sequences. The sequences of each clone were assembled by electropherogram analyses using the GAP 4 software of the Staden Package (Staden 1996).

Phylogenetic analyses

Amino acid and nucleotide sequences were aligned using the program MUSCLE (Edgar 2004). To illustrate the evolutionary relationship among *nik* and other *Drosophila* retrotranspsosons, a phylogenetic tree was constructed based on the amino acid alignment in the RT domain, using Neighbor-joining (NJ) method (p-distance model, pairwise deletion, 1000 bootstrap replicates) implemented on MEGA 4 (Tamura et al. 2007). Nucleotide sequences from *env*, *pol* and *gag* genes of *nik* obtained by *in silico* searches were used to infer *nik* phylogenies. Sequences obtained by cloning were included in the *env* phylogenies. Nucleotide phylogenetic trees were constructed by Maximum Parsimony (MP) (default parameters, 500 bootstrap replicates) and NJ method (Kimura-2-Parameter model, pairwise deletion, 1000 bootstrap replicates) using MEGA 4 software (Tamura et al. 2007).

Horizontal transfer test

High similarity between transposable element sequences from different species can indicate the occurrence of horizontal transfer (HT) or can be consequence of highly selective constraints. To investigate the possibility of *nik* HT events we analyzed the divergence only in the synonymous sites (dS). This approach offers a measure of neutral evolution in the absence of strong codon usage bias (CUB). If a TE was acquired via vertical transfer (VT), its dS value is expected to be similar to dS values of host genes. Alternatively, if the TE was horizontally transmitted to the host genome, the TE dS should be significantly lower than host genes dS (Silva and Kidwell, 2000; Ludwig et al. 2008; Bartolomé et al. 2009).

Synonymous divergence values for pairwise comparisons of a sample of 10,150 nuclear genes from *D. yakuba, D. simulans* and *D. melanogaster* (Begun et al. 2007) are nearly normally distributed. In this work, we used the information about dS distribution presented by Bartolomé et al. (2009). These authors estimated the 2.5% and 97.5% quantiles of the dS distributions by bootstrap, (mean [2.5%-97.5% quantiles]): 0.126 [0.037-0.230], 0.303 [0.096-0.531] and

0.284 [0.083-0.505], for *D. melanogaster* versus *D. simulans*, *D. melanogaster* versus *D. yakuba* and *D. simulans* versus *D. yakuba* comparisons, respectively.

The *nik* dS values found in pairwise comparisons among these three species were compared to their distribution of dS. HT events were considered when the level *nik* dS was lower than the 2.5% quantile of the dS for the nuclear genes of the hosts. Codon alignment for *nik pol* gene was used to estimate the dS values using the Nei and Gojobori (1986) method assisted by the MEGA 4 (Tamura et al. 2007).

RESULTS AND DISCUSSION

Genome searches

Using the complete *nik* as query in the BLASTn, several hits were found in *D. melanogaster, D. simulans, D. sechellia, D. erecta* and *D. yakuba.* However, few copies of *nik* are complete (including LTRs and the three gene regions). Most of sequences can be considered degenerated copies and are present in TE-rich regions, indicating ancient loss of activity. Table 1 summarizes the results for each species and the figure 1 shows a schematic organization of selected copies.

In *D. melanogaster*, we found 15 significant hits. Only one complete copy was found in the X chromosome (Figure 1A), which is the canonical copy, originally described by Bowen and McDonald (2001). This copy is probably recently inserted since it has identical LTRs and TSDs. An interesting degenerated copy (melanogaster10) is shown in the figure 1B. It has identical LTRs and TSDs, but presents a big internal deletion that has eliminated *gag, pol* and part of *env* genes. Also, downstream of the 5’ LTR of melanogaster10, there is an undescribed region (≅ 850 bp) that is absent in the canonical copy, but is present in some copies from other species. Other sequences in *D. melanogaster* are incomplete and degenerated, probably reminiscent copies of old insertions.

D. sechellia showed several hits, but no complete *nik* copy was found in this genome. *D simulans* has only one complete copy of *nik* (Figure 1C) and five other hits are from degenerated copies. *D. erecta* has a copy with a 140-bp missing region in the 3’ LTR, but is probably a complete sequence (Figure 1D). Other 5 copies in *D. erecta* are degenerated ones. In *D. yakuba* two nearly complete copies of *nik* were found, however yakuba1 (Figure 1E) has a missing region in the *env* gene and yakuba2 (Figure 1F) has a small deletion in the 5'LTR. Three other sequences found in *D. yakuba* are incomplete and degenerated copies. The complete copies from *D. simulans, D. yakuba* and *D. erecta* have the undescribed region downstream the 5' LTR.

BLASTn does not show any significant hit in the remaining species, *D. ananassae*, *D. pseudoobscura, D. persimilis, D. willistoni, D. virilis* and *D. grimshawi.* Nevertheless, using tBLASTx, some hits were found in all these species. One or two of the first hits were used in a phylogenetic analysis to access the identity of these sequences.

Cloning and sequencing

To further define the distribution of *nik*, and infer its phylogenetic relationship, a survey by PCR amplification was performed in 65 Drosophilidae species (table S1 of supplementary material). Amplification products of expected length (550 bp) were obtained only for the species of the *melanogaster* subgroup (*D. melanogaster, D. simulans, D. teissieri, D. yakuba, D. erecta, D. sechellia, D. santomea* and *D. mauritiana*). Amplicons from *D. teissieri*, *D. santomea* and *D. mauritiana* were cloned and sequenced to include in the phylogenetic analyses.

Phylogenetic analyses

Nucleotide phylogenies help to understand the diversity and relationship of the *nik* sequences found in this study. Figures 2A, 2B and 2C show, respectively, *pol, env* and *gag* phylogenetic NJ trees. We performed these three separately analyses in order to use all sequences found *in silico* that were often incomplete. The sequences of *nik* can be clearly divided into 4 groups that are consistent in the three phylogenies. The MP trees presented the same basic topology and the bootstrap values of main clades are shown in brackets in the NJ trees. The only major difference was found in the *env* phylogeny that groups the sequence yakuba6 in the group II instead group I as in the NJ tree.

The group I is composed by degenerated sequences from *D. melanogaster, D. sechellia* and *D. simulans* and is closely related to the group II. The group II includes the canonical *nik* sequence and is present in all analyzed species of *D. melanogaster* subgroup, with the exception of *D. sechellia*. Clone sequences from *D. santomea, D. mauritiana* and *D. teissieri* are grouped in this clade. This distribution of canonical *nik* suggests that it may be an old retrotransposon present in the ancestor of *melanogaster* subgroup and was probably lost in *D. sechellia*. However, these sequences show high similarity as can be seen by the very short branch length. The complete *nik* sequences obtained from *D. melanogaster*, *D. simulans* and *D. yakuba* show especially high similarity (99.7%). Thus, it is possible that HT events have occurred recently among these species (see below).

The group III is composed by sequences from *D. erecta, D. yakuba, D. simulans* and *D. sechellia*. Group IV has only sequences from the sister species *D. simulans and D. melanogaster*. Both groups have more divergent sequences that are much degenerated. We need to consider the occurrence of ancestral polymorphism and stochastic loss to elucidate some of the relationships in these groups.

We use the RT domain to better understand the position of *nik* and related sequences in the *Drosophila* retrotransposons phylogeny (figure 3). Representative sequences from *Ty3-gypsy, Ty1-copia* and *Bel-Pao* groups were included in this tree along with some *nik* sequences found *in silico*. The phylogeny confirms that *nik* belongs undoubtedly to the *Ty3-gypsy* clade (Bowen and McDonald 2001), more specifically to the *gypsy-like* lineage. Some sequences found in *D. ananassae* (ananassae1), *D. pseudoobscura* (pseudoobscura1), *D. mojavensis* (mojavensis2) and *D. willistoni* (willistoni2) are more related to the retrotransposons *Accord*, *ZAM* and *Tirant* and the sequence from *D. grimshawi* (grimshawi1) is more related to the retrotransposon *Quasimodo.* Sequences found in *D. virilis* correspond to the described retrotransposon *TV1.* There are three sequences outside *melanogaster* subgroup that are more closely related to *nik* than to other retrotransposons, persimilis1 from *D. persimilis,* willistoni1 from *D. willistoni* and mojavensis1 from *D. mojavensis*. These sequences were better analized and we detect they are degenerated copies with no LTRs. These sequences can be reminiscent of *nik* family in these species suggesting a very ancient origin of *nik* prior to the split of *Drosophila* and *Drosophila* subgenera.

High similarity among complete *nik* sequences

The complete sequences from *D. yakuba* (yakuba1), *D. simulans* (simulans1) and *D. melanogaster* (melanogaster1) present very high similarity (overall mean 99. 7 %). These species are very closely related as can be seen in the figure 2D, but even in this case, a higher level of divergence would be expected if these *nik* copies were transmitted vertically by the ancestor of *melanogaster* subgroup.

To test HT hypothesis we compare the dS values of *nik pol* gene, obtained for pairwise comparisons, with those expected assuming vertical transmission. The dS values obtained (0.0029 for *D. melanogaster* versus *D. simulans*, 0.0029 from *D. melanogaster* versus *D. yakuba* and 0.0000 for *D. simulans* versus *D. yakuba*) are significantly lower than expected (see material and methods) suggesting that these sequences were probably involved in horizontal transfer among these species.

A brief analysis of *env* sequences from the other species reveals that all sequences from group II, comprising the canonical *nik,* present little total divergence (0.03%) and present dS values much lower than those found for *alpha metildopa* (*amd*) gene, a commonly used gene for HT tests (Ludwig et al. 2008; Vidal et al. 2008) (data not shown). We also discard a possible high codon usage bias on *nik* that could lead to lower dS values. Thus, several HT events would be necessary to explain the high conservation of *nik* among species. An alternative explanation, not mutually exclusive, is the occurrence of introgression events. As can be seen in the figure 2D, the speciation in the *melanogaster* subgroup starts around 12 million years ago and, for a long time, gene flow could have occurred among these species.

However, other phenomena may be held responsible for the distortions in the measurement of dS; eg, the sites required in splicing mechanisms or those involved in RNA secondary structure and other aspects linked to functionality of RNAs (Ngandu et al 2008; Xing and Lee 2006). Thus, the conservation of *nik* dS could be result of some kind of selection to maintain the nucleotide sequences for an unknown function.

nik **gene organization**

Full length copies of the endogenous retrovirus *nik* have 430-bp LTRs flanking the *gag, pol* and *env* genes. The undescribed region found after the 5'LTR in some complete copies has no similarity with any known sequence in the genbank, as revealed by BLAST searches. It is a non-coding region and it was probably deleted in the melanogaster1 copy from *D. melanogaster*.

ORF Finder reveals the presence of three ORFs in the canonical *nik* from *D. melanogaster*: *gag* (frame +2; 385 amino acids), *pol* (frame +1; 1062 amino acids) and *env* (frame +3; 412 amino acids). However, the predicted Pol protein does not have the first domain, PR.

For most retroelements, ribosomal frameshifting is a common strategy employed to express Gag-Pol. At a particular step in the cycle of translational elongation the ribosome shifts its reading frame from the one it initiated translating into a new reading frame (Farabaugh 1996). The presence of an overlapping region between *gag* and *pol* genes with ribosomal frameshifting could generate a complete Pol protein of *nik*. We then aligned and compared the possible *gag-pol* overlapping region of *nik* with those from the related retrotransposons *17.6, 297, Tom* and *Idefix* that have the putative frameshift site already identified (Figure 4) (Gao et al. 2003). We observed that, *nik* has a common structural motif of -1 frameshift, which is the most prevalent form of retroelements gene organization (Farabaugh 1996; Gao et al. 2003). The sites consist of a heptameric sequence of the form X-XXY-YYZ (Weiss et al. 1989). Considering the putative frameshift site of *nik*, the predicted Pol polyprotein has 1173 amino acids including the PR domain. The production of Gag-Pol is likely important for packing the Pol products within the viral particles (Farabaugh 1996; Gao et al. 2003).

***nik* regulation**

In some eukaryotes, small RNA molecules, called piRNAs, are responsible to defend the genome against transposable elements. In *Drosophila*, distinct heterochromatic loci (clusters) are the source of primary antisense piRNAs, which then target a large number of transposable elements that are dispersed trough-out the genome (Brennecke et al. 2007). The *flamenco*, originally identified as a locus controlling the *gypsy* element (Pelisson et al. 1994), is a major piRNA cluster and consists of a large number of truncated or defective retrotransposons (Brennecke et al. 2007). Recently, Malone et al. (2009) showed that there are distinct piRNA pathways acting in the germline and somatic tissues of ovary. The *flamenco* locus was appointed as the main piRNA cluster acting in the soma, to target elements *gypsy-like*. In the germline, a variety of clusters collaborate to control a broad range of elements.

Aiming to find some information about the regulation of *nik,* we investigated which regions of *nik* are present in the *flamenco* locus in three species, *D. melanogaster*, *D. yakuba* and *D. erecta*. As we can observe in the figure 5, *D. melanogaster* and *D. erecta* have only a very small region of *nik* in the *flamenco* cluster, while *D. yakuba* has several regions comprising the three genes. This observation suggests that *flamenco* locus can enclose important differences in the content of retrotransposons among species, although it was present in the common ancestor of *melanogaster* subgroup species. One might suggest that *flamenco* has been shaped by its coevolution with active retrotransposons that need to be controlled and with other piRNA clusters that can contribute to the regulation work. This coevolution process must ensure efficient defense against retrotransposons.

Despite we found a very small region of *nik* in the *D*. *melanogaster flamenco* locus, a high density of antisense piRNAs, from somatic cells ovary, is found matching the entire *nik* sequence (Malone et al. 2009). This could indicate that other loci, instead *flamenco*, may be involved in the production of *nik* piRNAs in the *D. melanogaster* soma, although the participation of *flamenco* in the *D.*

melanogaster nik regulation is suggested by an overall decrease of *nik* piRNA levels in *flamenco* mutants (Malone et al. 2009). How *flamenco* could affect *nik* piRNAs production is an open question.

To find another cluster putatively involved in the *nik* piRNAs production, we compared the location of *nik* homologous sequences that we found in this work with the chromosome location of 142 piRNA clusters identified in *D. melanogaster* (Brennecke et al. 2007). The piRNA cluster 19, located in the chromosome X: 11089529-11095706, corresponds to the complete *nik* copy (melanogaster1). It's not surprisingly that *nik* piRNAs would map in the complete copy, since it is the silencing target. However, most of piRNAs mapping in this locus (50 from 59) does not map in other region in the genome (Brennecke et al. 2007) suggesting that the complete *nik* copy is a primary source of those piRNAs. Alternatively, the cluster that provides *nik* piRNAs can be located in a sequencing gap region of the *D. melanogaster* genome.

Concluding remarks

In this work, we present a detailed analysis of the evolution of *nik*, a *Drosophila* endogenous retrovirus. We found that *nik* could be present in the ancestor of *Drosophila* and *Drosophila* subgenera (as suggested by the presence of *nik* homologous sequences in *D. persimilis, D. willistoni* and *D. mojavensis*) and was probably lost in major of species. The presence of 4 distinct groups of *nik* sequences, from *melanogaster* subgroup species, implies that some diversification of this endogenous retrovirus have occurred prior to the *melanogaster* subgroup speciation. Different evolutionary process, as independent assortment of copies, stochastic lost, introgression, and horizontal transfer, help to explain the current distribution of *nik*. The presence of complete *nik* sequences with identical LTRs indicates recent activity of this endogenous retrovirus in the *Drosophila* genomes.

REFERENCES

Bartolomé C, Bello X and Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. Genome Biol 10:R22.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E and Langley CH (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310.

Bowen NJ and McDonald JF (2001) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11:1527-1540.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R and Hannon GJ (2007) Discrete small RNA-generating *loci* as master regulators of transposon activity in *Drosophila*. Cell 128:1089-1103.

Coffin JM, Hughes SH and Varmus HE (1997) Retroviruses. Cold Spring Harbor Laboratory Press, New Yor, pp 843.

Desset S, Meignin C, Dastugue B and Vaury C (2003) COM, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. Genetics 164:501-509.

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

Eickbush TH and Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221-234.

Farabaugh PJ (1996) Programmed translational frameshifting. Microbiol Rev 60:103-134.

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

Grumbling G and Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34:D484-8.

Havecker ER, Gao X and Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.

Herédia F, Loreto EL, Valente VL (2004) Complex Evolution of *gypsy* in Drosophilid Species. Mol Biol Evol 21(10):1831-1842.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, Ashburner M and Celniker SE (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3:RESEARCH0084.

Kapitonov VV and Jurka J (2000) GYPSY5_I. Direct Submission to Repbase Update (SEP-2000).

Kapitonov VV and Jurka J (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A 100:6569-6574.

Kim A, Terzian C, Santamaria P, Pélisson A, Purd'homme N and Bucheton A (1994) Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc Natl Acad Sci USA 91:1285-1289.

Kohany O, Gentles AJ, Hankus L and Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B and Vaury C (2000) Life cycle of an endogenous retrovirus, *ZAM*, in *Drosophila melanogaster*. J Virol 74:10658-10669.

Lewis RL, Beckenbach AT and Mooers AØ (2005) The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on *COI* and *COII* sequences and a Bayesian estimate of phylogeny. Mol Phylogenet Evol 37:15-24.

Llorens JV, Clark JB, Martínez-Garay I, Soriano S, de Frutos R and Martínez-Sebastián MJ (2008) *Gypsy* endogenous retrovirus maintains potential infectivity in several species of Drosophilids. BMC Evol Biol 8:302.

Ludwig A, Loreto EL (2007) Evolutionary pattern of the gtwin retrotransposon in the *Drosophila* melanogaster subgroup. Genetica 130(2):161-8.

Ludwig A, Valente VL and Loreto EL (2008) Multiple invasions of *Errantivirus* in the genus *Drosophila*. Insect Mol Biol 17:113-124.

Malik HS, Henikoff S and Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307-1318.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R and Hannon GJ (2009) Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. Cell 137:522-535.

Mejlumian L, Pélisson A, Bucheton A and Terzian C (2002) Comparative and functional studies of *Drosophila* species invasion by the *gypsy* endogenous retrovirus. Genetics 160:201-209.

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

Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D and Seoighe C (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J 5:160.

Pélisson A, Song SU, Prud'homme N, Smith PA, Bucheton A and Corces VG (1994) *Gypsy* transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the *Drosophila flamenco* gene. EMBO J 13:4401-4411.

Prud'homme N, Gans M, Masson M, Terzian C and Bucheton A (1995) *Flamenco*, a gene controlling the *gypsy* retrovirus of *Drosophila melanogaster*. Genetics 139:697-711.

Silva JC and Kidwell MG (2000) Horizontal transfer and selection in the evolution of *P* elements. Mol Biol Evol 17:1542-1557.

Song SU, Gerasimova T, Kurkulos M, Boeke JD and Corces VG (1994) An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev 8:2046-2057.

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

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599.

Tamura K, Subramanian S and Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

Terzian C, Ferraz C, Demaille J and Bucheton A (2000) Evolution of the *Gypsy* endogenous retrovirus in the *Drosophila melanogaster* subgroup. Mol Biol Evol 17:908-914.

Vidal NM, Ludwig A and Loreto ELS (2009) Evolution of *Tom, 297, 17.6* and *rover* retrotransposons in Drosophilidae species. Mol Gen Genomics 282:351-362.

Weiss RB, Dunn DM, Shuh M, Atkins JF and Gesteland RF (1989) *E. coli* ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. New Biol 1:159-169.

Xing Y and Lee C (2006) Can RNA selection pressure distort the measurement of Ka/Ks? Gene 29:1-5.

TABLES

Table 1: *nik* sequences identified in this study

Species	Position	Sequence name	Total length (bp)	LTR length (bp)	LTR similarity*	TSD
D. melanogaster	gnl|dmel|X: 11088517-11095880	melanogaster1	7364	430	1.00	CCAG
	gnl|dmel|X: 21829034-21831646	melanogaster2	2623	-	-	-
	gnl|dmel|U:263419-267959	melanogaster3	4541	-	-	-
	gnl|dmel|U:256391-261910	melanogaster4	5520	-	-	-
	gnl|dmel|U:6060871-6064621	melanogaster5	3751	-	-	-
	gnl|dmel|U:1745754-1750147	melanogaster6	4394	-	-	-
	gnl|dmel|U:3305941-3309659	melanogaster7	3719	-	-	-
	gnl|dmel|U:3238960-3243189	melanogaster8	4230	-	-	-
	gnl|dmel|U:3243622-3247851	melanogaster9	4230	-	-	-
	gnl|dmel|3L:15288923-15291708	melanogaster10	2786	430	1.00	CGCG
	gnl|dmel|3LHet:764699-769082	melanogaster11	4384	-	-	-
	gnl|dmel|3RHet:1340664-1351729	melanogaster12	5812	-	-	-
	gnl|dmel|2RHet:1763082-1765784	melanogaster13	2703	-	-	-
	gnl|dmel|2RHet:1664479-1667377	melanogaster14	2889	-	-	-
	gnl|dmel|2RHet:1794061-1798454	melanogaster15	4394	-	-	-
D. simulans	gnl|dsim|X:15883937-15892224	simulans1	7333	430	1.00	CGCG
	gnl|dsim|2R:1155165-1160181	simulans2	5017	-	-	-
	gnl|dsim|chr2h_Mrandom_024:171352-178960	simulans3	5897	-	-	-
	gnl|dsim|chr2h_Mrandom_038:170114-176461	simulans4	5570	-	-	-
	gnl|dsim|chrU_M_3474:2192-5772	simulans5	3191	-	-	-
	gnl|dsim|chrU_M_1770:19737-21186	simulans6	1450	-	-	-
D. sechellia	gnl|dsec|scaffold_126:41105-46429	sechellia1	5325	-	-	-
	gnl|dsec|scaffold_238:2516-8765	sechellia2	6260	-	-	-
	gnl|dsec|scaffold_156:44111-49298	sechellia3	5188	-	-	-
	gnl|dsec|scaffold_91:70488-75092	sechellia4	4605	-	-	-
	gnl|dsec|scaffold_86:112773-122552	sechellia5	5126	-	-	-
	gnl|dsec|scaffold_62:13513-19293	sechellia6	5396	-	-	-
	gnl|dsec|scaffold_482:784-6673	sechellia7	5890	-	-	-
	gnl|dsec|scaffold_66:18196-23982	sechellia8	5787	-	-	-
	gnl|dsec|scaffold_232:6957-13165	sechellia9	6209	-	-	-
	gnl|dsec|scaffold_401:8291-13570	sechellia10	5280	-	-	-
D. yakuba	gnl|dyak|v2_chrX_random_022:104496-112408	yakuba1	7912	430	0.997	CGCG
	gnl|dyak|v2_chr3h_random_002:547483-554270	yakuba2	6788	396/430	0.997	CGCG
	gnl|dyak|v2_chrX_random_023:1-1867	yakuba3	1867	-	-	-
	gnl|dyak|v2_chrX_random_024:1-1701	yakuba4	1701	-	-	-
	gnl|dyak|3R:262981-272338	yakuba5	6515	-	-	-
D. erecta	gnl|dere|scaffold_4772:41940-49920	erecta1	7189	430/287	0.996	-
	gnl|dere|scaffold_4929:24405266-24412515	erecta2	7250	-	-	-
	gnl|dere|scaffold_4694:83555-87174	erecta3	3620	-	-	-
	gnl|dere|scaffold_4694:88685-92296	erecta4	3612	-	-	-
	gnl|dere|scaffold_4694:92686-95843	erecta5	3158	-	-	-
	gnl|dere|scaffold_4714:115053-119695	erecta6	4643	-	-	-

* LTR similarity was calculated excluding deleted regions

FIGURES

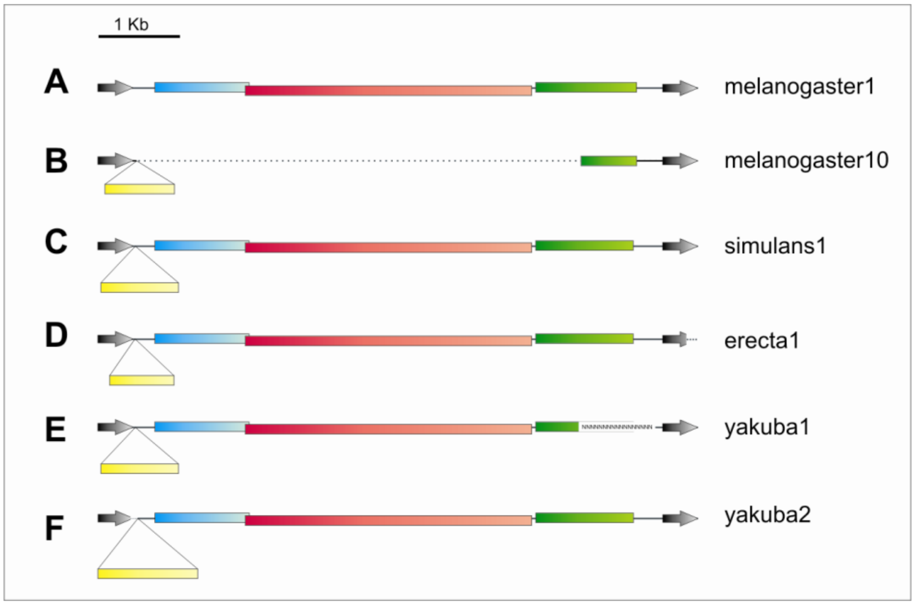


Figure 1: Schematic representation of *nik* copies found in different genomes. Gray arrows – LTRs; blue rectangle – *gag* gene; red rectangle – *pol* gene; green rectangle – *env* gene; yellow rectangle – undescribed region; dashed line – deleted region; N – missing region.

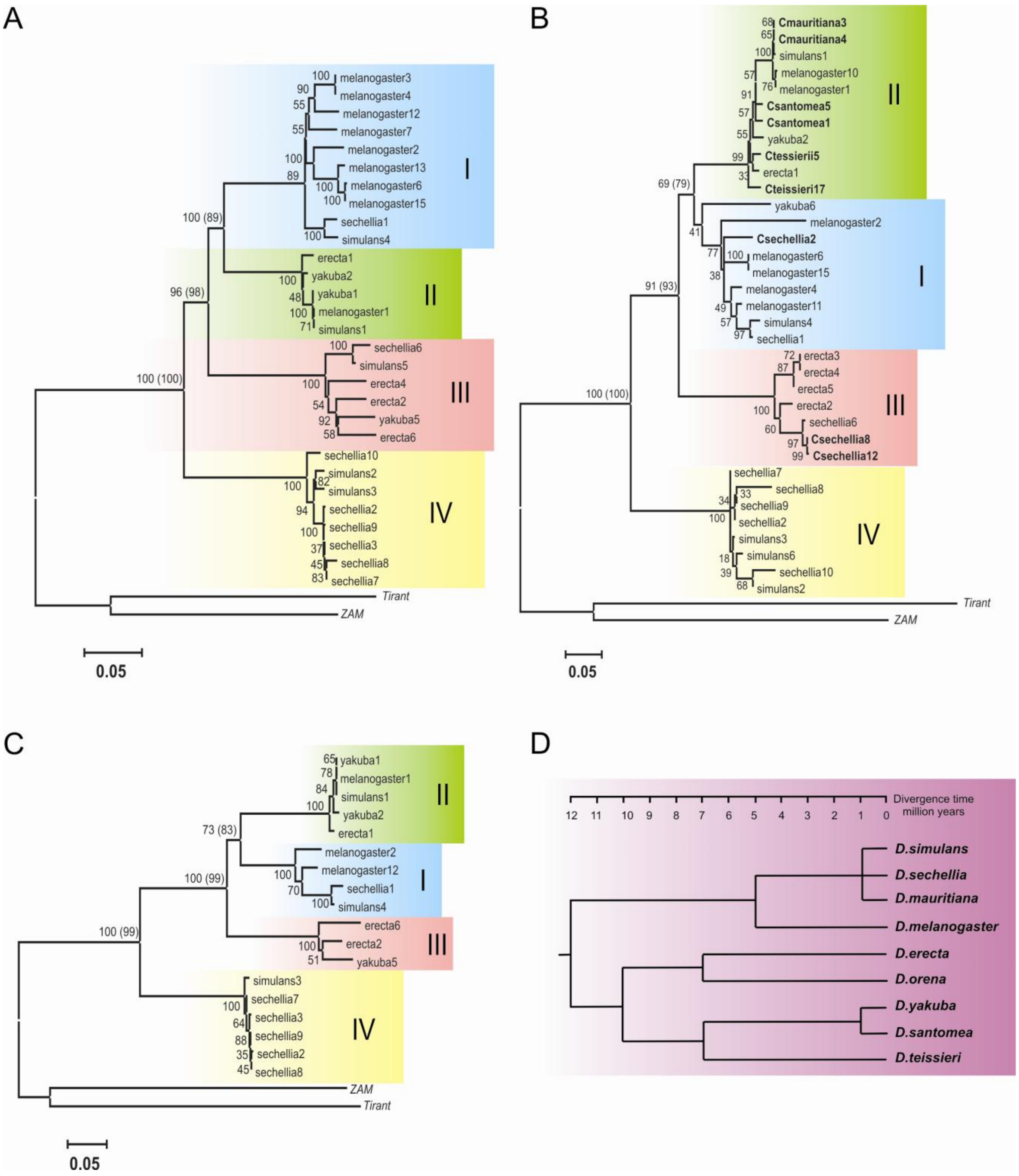


Figure 2: Phylogeny of *nik* (A, B and C) and host species (D). NJ trees of *nik* were constructed based on regions of *pol* (A), *gag* (B) or *env* genes (C) and were divided into four main clades, I, II, III and IV. NJ bootstrap values are shown in the nodes and the MP bootstrap values for the main nodes are shown into brackets. In the *env* tree (B), sequences originated by cloning are showed in bold and with the letter C followed by the name of species and the number of the clone. Sequences from the *D. melanogaster* endogenous retroviruses *ZAM* and *Tirant* were used as outgroups. In (D) we show a schematic representation of the evolutionary relationships of *nik* host species, the *melanogaster* subgroup (based on Lewis et al. 2005). The estimated divergence time of species are shown in the time scale (based on Tamura et al. 2004).

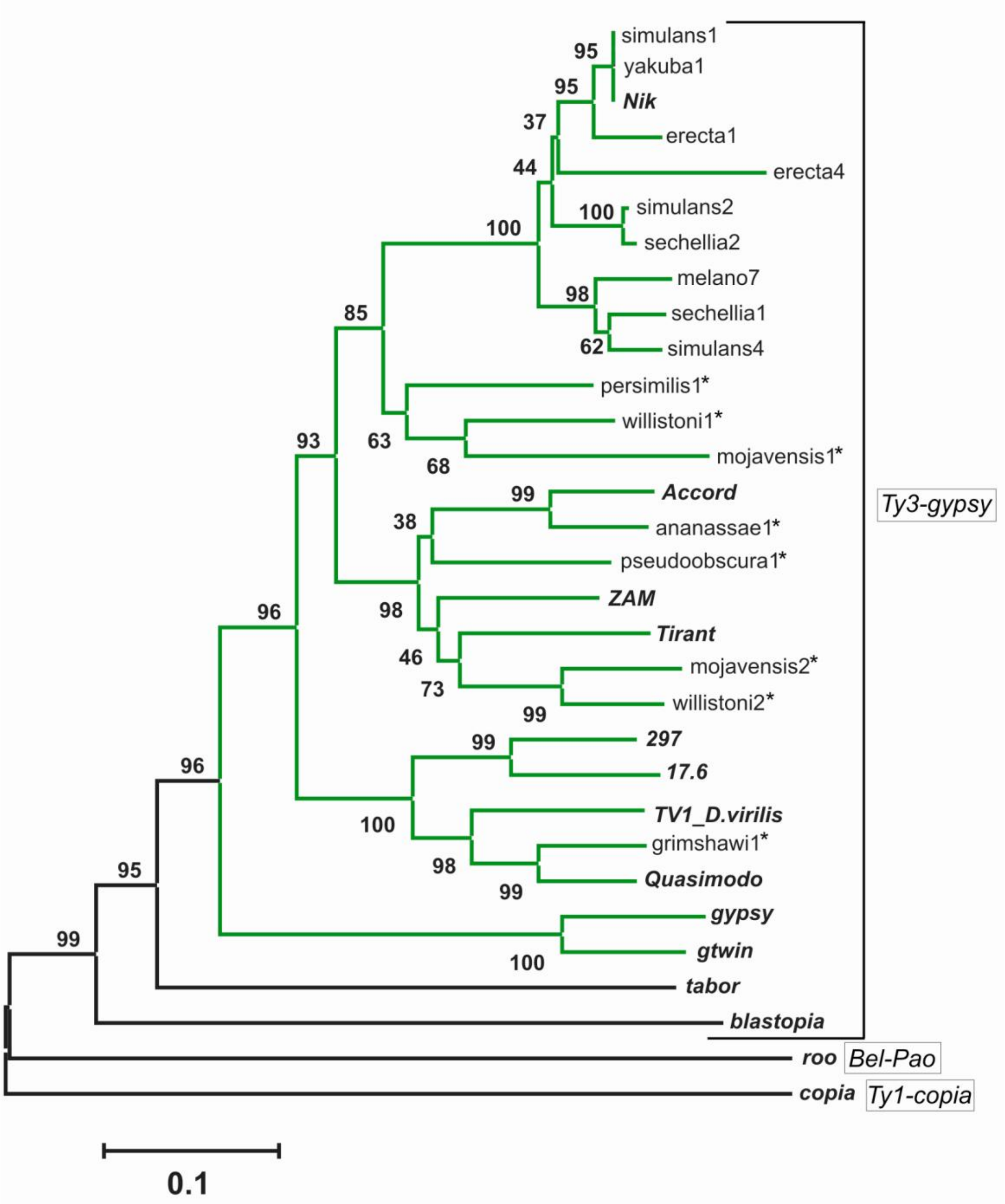


Figure 3: NJ phylogeny of selected representatives from the *Ty3-gypsy*, *Ty1-copia* and *Bel-Pao* clades of retrotransposons and *nik* sequences. The phylogeny is based on the reverse transcriptase domain. The branches in green represent the *gypsy-like* lineage, a monophyletic group of insect endogenous retroviruses with an *env* gene. The sequences found by tBLASTx are marked with asterisk (*) and are supposed to have *env* gene by the position in the phylogeny although complete copies were no analyzed.

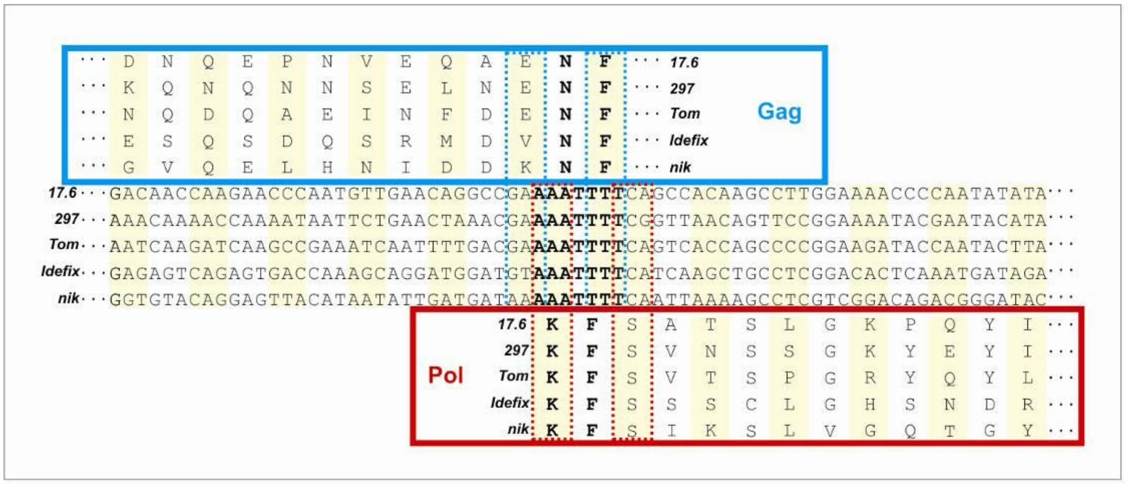

··· D N Q E P N V E Q A E N F ··· 17.6
··· K Q N Q N N S E L N E N F ··· 297
··· N Q D Q A E I N F D E N F ··· Tom
··· E S Q S D Q S R M D V N F ··· Idefix
··· G V Q E L H N I D D K N F ··· nik
Gag
17.6 ··· GACAACCAAGAACCCAATGTTGAACAGGCCGAAAATTTTCAGCCACAAGCCTTGGAAAACCCCAATATATA ···
297 ··· AAACAAAACCAAAATAATTCTGAACTAAACGAAAATTTTCGGTTAACAGTTCCGGAAAATACGAATACATA ···
Tom ··· AATCAAGATCAAGCCGAAATCAATTTTGACGAAAATTTTCAGTCACCAGCCCCGGAAGATACCAATACTTA ···
Idefix ··· GAGAGTCAGAGTGACCAAAGCAGGATGGATGTAAATTTTCATCAAGCTGCCTCGGACACTCAAATGATAGA ···
nik ··· GGTGTACAGGAGTTACATAATATTGATGATAAAAATTTTCAATTAAAAGCCTCGTCGGACAGACGGGATAC ···
Pol
17.6 K F S A T S L G K P Q Y I ···
297 K F S V N S S G K Y E Y I ···
Tom K F S V T S P G R Y Q Y L ···
Idefix K F S S S C L G H S N D R ···
nik K F S I K S L V G Q T G Y ···

Figure 4: Codon and amino acid alignments of *nik* and other retrotransposons with similar putative frameshifting, in the overlapping region between *gag* and *pol*. The nucleotides on the frameshift site are in bold conform the consensus X XXY YYZ. In these case nucleotides Z are equal to Y.

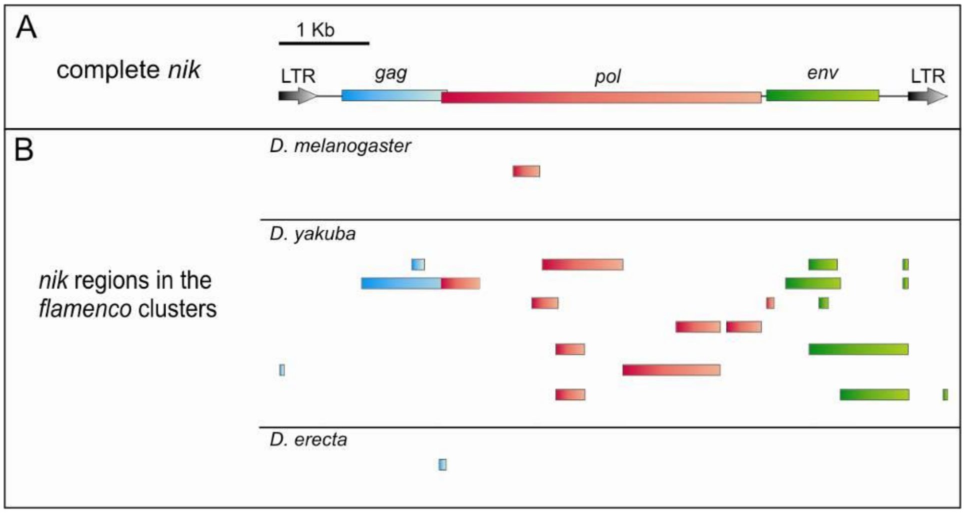


Figure 5: (A) Schematic organization of *nik.* (B) Comparison of *nik* regions found in the flamenco clusters in *D. melanogaster*, *D. yakuba* and *D. erecta*.

SUPLEMENTARY MATERIAL

Table S1: Drosophilidae species used in the PCR screenings.

Genus	Section	Group	Species
Drosophila	*quinaria-tripunctata*	*guarani*	*D. ornatifrons*
			D. subbadia
			D. guaru
		guaramunu	*D. griseolineata*
			D. maculifrons
		tripunctata	*D.nappae*
			D. paraguayensis
			D. crocina
			D. paramediostriata
			D. tripunctata
			D. mediodifusa
			D. bandeirantorum
			D. mediopictoides
		cardini	*D. cardini*
			D. cardinoides
			D. neocardini
			D. polymorpha
			D. procardinoides
			D. arawacana
		pallidipennis	*D. pallidipennis*
		calloptera	*D. ornatipennis*
		immigrans	*D. immigrans*
		funebris	*D. funebris*
	virilis- repleta	*mesophragmatica*	*D. gasici*
			D. brncici
			D. gaucha
			D. pavani
		repleta	*D. hydei*
			D. eohydei
			D. mercatorum
		anulimana	*D. anulimana*
		canalinea	*D. canalinea*
		flavopilosa	*D. cestri*
			D. incompta
		virilis	*D. virilis*
		robusta	*D. robusta*
		melanogaster	*D.melanogaster*
			D. simulans
			D. sechellia
			D. mauritiana
			D. teissieri
			D. santomea
			D. erecta
			D. yakuba
			D. kikkawai
			D. ananassae
			D. malerkotliana
		obscura	*D. pseudoobscura*
		saltans	*D. prosaltans*
			D. saltans
			D. neoliptica

		D. sturtevanti
	willistoni	*D.sucinea*
		D. nebulosa
		D. paulistorum
		D. willistoni
		D. equinoxialis
		D. insularis
		D. tropicalis
		D. capricorni
		D. busckii
Zaprionus		*Z. indianus*
		Z. tuberculatus
Scaptodrosophila		*S. latifasciaeformis*
		S. lebanonensis

Capítulo 4

Mar, a MITE family of *hAT* transposons in *Drosophila*

**Adriana Ludwig[1]*, Maríndia Deprá[1]*, Vera L. S. Valente[1],
Elgion L. S. Loreto[1,2]**

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

[2] Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), CEP 97105-900, Santa Maria, Rio Grande do Sul, Brazil; phone 55-55-32208912; email – elgion.loreto@pq.cnpq.br.

*These authors contributed equally to this project and should be considered co-first authors.

Manuscrito em preparação para ser enviado à revista Molecular Genetics and Genomics.

ABSTRACT

Miniature inverted-repeat transposable elements (MITEs) are short nonautonomous DNAs elements, flanked by subterminal or terminal inverted repeats (TIRs). MITEs were recognized as important components of plant genomes where they can attain extremely high copy numbers. They are found also in several animal genomes, including mosquitoes, fish and humans. In *Drosophila*, few families were described up to now. Here, we investigated the distribution and evolution of the element *Mar*, a MITE, in Drosophilidae species. The *Mar* distribution is restrict to the *willistoni* subgroup of *Drosophila* and the phylogeny supports the view that the origin of this element might have occurred prior to diversification of these species. Most of the *Mar* copies in *D. willistoni* present conserved TSDs (target duplication site) and TIRs, indicating recent mobilization of these sequences. Additionally, many of these copies were found in or near genes suggesting that *Mar* can promote important genetic variability within and between species.

Keywords: Miniature inverted-repeat transposable elements, MITE, *Mar* element, transposon, *Drosophila*

INTRODUCTION

Transposable elements (TEs) are discrete segments of DNA that are distinguished by their ability to move and replicate within genomes (Kidwell and Lisch 2001). TE-derived sequences are the most abundant component of practically all eukaryotic genomes. An increasing amount of evidences have shown that TEs can play an important role in driving the evolution and genome complexity (Kazazian 2004; Thornburg et al. 2006; Feschotte and Pritham 2007; Deprá et al. 2009).

TEs can be divided into two classes, based on their mechanism of transposition: class I comprises retrotransposons that transpose by a RNA intermediate, and class II comprises transposons that move through DNA intermediate (Wicker at al. 2007). Depending on their ability to direct their own transposition, each class of TEs can contain two types: autonomous and non-autonomous copies. Autonomous TEs encode for the proteins required for their transposition, while non-autonomous TEs can be mobilized *in trans* by the enzymes produced by autonomous elements (Capy et al. 1998).

Within the classe II transposons there is a special group of non-autonomous sequences, miniature inverted-repeat transposable elements (MITEs), which can be present in high copy numbers in the genomes. They are characterized by short sequences with no coding capacity, flanked by subterminal or terminal inverted repeats (TIRs) and short direct repeats caused by target site duplication (TSDs). They are probably originated from a subset of autonomous DNA transposons (Bureau and Wessler 1992; Jiang et al. 2003; Quesneville et al. 2006; Ortiz and Loreto 2008). MITEs have no internal homology to their parental autonomous transposons and often include non-homologous AT-rich sequences in their internal regions. Besides of the first discovery in plants, MITEs have also been found in several animal genomes, including *Caenorhabditis elegans*, *Drosophila*, mosquitoes, fish and humans (Feschotte et al. 2002; Gonzáles and Petrov 2009).

In *Drosophila*, there are few described MITE families. Two of them, were discovered in *D. willistoni*, *Vege* and *Mar,* by Holyoake and Kidwell (2003). These elements are short, 884 bp and 610 bp respectively, are AT-rich and have 8-bp

TSDs. They were described to have perfect TIRs of 12-bp for *Vege* and 11-bp for *Mar*. TBLASTn and BLASTx analysis indicate that both elements have neither coding capacity nor significant sequence homology to known published sequences. As MITEs have been grouped into superfamilies based on their TIRs and TSDs length in association with the superfamilies of transposases, *Vege* and *Mar* were classified as MITEs from the *hAT* superfamily. Thus, the precursor elements of *Mar* and *Vege* are probably autonomous elements from *hAT* superfamily; however they were not identified (Holyoake and Kidwell 2003). The *hAT* superfamily is widely distributed in multicellular organisms, including plants, animals and fungi (Rubin et al. 2001). Members of this superfamily have TSDs of 8 bp, relatively short TIRs of 5–27 bp and overall lengths of less than 4 kb (Wicker et al. 2007).

Plant genomes, in particular rice, are the model organism to study MITEs (Gonzáles and Petrov 2009). Little is known about MITEs in *Drosophila*. Here, we investigated the distribution and evolution of the element *Mar* in Drosophilidae species, as well as, we characterized the *Mar* copies present in *D. willistoni* genome. We show that the *Mar* element is restricted to the *willistoni* subgroup species. We propose that the *Mar* origin have occurred after the separation of *willistoni* and *bocainensis* subgroups, which is consistent with their distribution. In *D. willistoni* we found evidences of recent mobilization and burst of transposition.

MATERIAL AND METHODS

in silico searches

Searches for sequences homologous to the element *Mar* were conducted in the *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. virilis, D. mojavensis, D. grimshawi* and *D. willistoni* genomes using the BLAST tool (http://flybase.bio.indiana.edu/blast/; Grumbling and Strelets 2006). The complete *Mar* sequence (access number: AF518731.1) was used as query. The *Mar* sequences from *D. willistoni* were analyzed for the presence of conserved TIRs and TSDs by visual inspection of alignments. WebLogo was used to analysis of

TSDs (Crooks et al. 2004). Local BlastN were performed against different sequences dataset of *D. willistoni* genome (CDS, intron and gene extended 2000) in order to obtain a view of the *Mar* insertions in gene regions.

PCR and Dot blot Screening

In order to analyze the distribution of the element *Mar* we conducted screening by PCR and Dot blot. A total of 65 *Drosophila* species were screened (Table 1). DNA was extracted from 30 fresh adult flies according to Sassi et al. (2005). For PCR reactions, two primers were designed for amplifying a fragment of approximately 450 bp of the *Mar* element, MarF 5' CGCGAATCGTATGTGAA 3' and MarR 5' CGATGTGAGCACGAAGTACA 3'. PCR reactions (50 µl) were performed using conditions as follows: 50 ng template DNA, 20 pM of each primer, 2.5 mM $MgCl_2$, and 1 U *Taq* DNA polymerase. Amplification was implemented with denaturing at 92 °C for 2 min, 30 cycles of denaturing at 92°C for 45 s, annealing at 55 °C for 50 s, and extension at 72 °C for 1 min, followed by extension at 72 °C for 5 min.

For dot blot hybridizations, samples of denatured DNA (1 µg) were transferred onto a nylon membrane (Hybond-N+; GE Healthcare). AlkPhos Direct Labelling and Detection System with CDP-Star kit (GE Healthcare) were used to label and detected nucleic acids following the manufacturer’s instructions. *Mar* element from *D. willistoni* was used as probe.

DNA Cloning and Sequencing

The amplified samples were run on 0.8% agarose gel and the bands excised using GFX Purification Kit (GE Healthcare) and cloned with TOPO-TA cloning vector (Invitrogen). The cloned PCR products were sequenced using the universal primers, M13 forward and M13 reverse, in Megabace 500 sequencer. The dideoxy chain-termination reaction was implemented using the DYEnamicET kit (GE Healthcare). Both DNA strands were sequenced at least twice or until to obtain very reliable sequences. The sequences of each clone were assembled by electropherogram analyses using the GAP 4 software of the Staden Package (Staden 1996).

Sequence Alignment and Phylogenetic Analysis

Sequences were aligned using Muscle tool (Robert 2004) with default parameter values. Phylogenetic trees were constructed by Neighbor-Joining method (Saitou and Nei 1987) with the Kimura-2-Parameter nucleotide substitution model, and 1000-replication bootstrap, using MEGA 4 software (Tamura et al. 2007).

RESULTS AND DISCUSSION

Element *Mar* is restricted to *willistoni* subgroup species

As expected, the element *Mar* was found in the *D. willistoni* genome by *in silico* searches. In the other 11 genomes investigated no homologous sequences to *Mar* were found.

To increase the examination of *Mar* distribution we used PCR and Dot blot strategies in a large number of Drosophilidae species that belong to different *Drosophila* groups (Table 1). PCR results showed amplification only in the species from *willistoni* subgroup, *D. willistoni*, *D. paulistorum, D. equinoxialis, D. insularis* and *D. tropicalis.* The fragment amplified in *D. tropicalis* was larger than expected (around 3,000 bp) suggesting the possibility of finding a full-length transposon. The fragments length varies from roughly 270 bp to 450 bp and most of the length variation can be seen between species. Dot blot result (Figure S1 of supplementary material and Figure 1) corroborated PCR results showing positive signal only in the *willistoni* subgroup species. Also, *D. tropicalis* that present a very large sequence was weaker signal than other *willistoni* subgroup species. Species from *bocainensis* subgroup (*willistoni* group) present a very week signal, what may indicate the presence of sequences related to *Mar* although with high divergence.

All cloned sequences and major of those obtained by *in silico* searches were used in the phylogenetic analysis to understand the evolutionary dynamics of *Mar* in the *willistoni* group species. The figure 2 shows the Neighbor-joining tree obtained for *Mar,* that can be compared with the relationships among *willistoni* group species (Figure 1) established recently (Robe et al. 2010). We can divide the phylogeny into 3 groups, A, B and C. The groups A and C are composed only by sequences from *D. willistoni*. The group C has less and more divergent sequences while the group A represents a clear burst of transposition in *D. willistoni.* Group A is closely related to the group B that has *Mar* sequences from other species. There is very little clustering by species in this group. It could be indicative of horizontal transfer between species, a not rare process in TE evolution (Loreto et al. 2008). However, the species involved are very closely related and some levels of incongruence were found between different phylogenies of *willistoni* subgroup, in which saturation, introgression and maybe ancestral polymorphisms incompletely sorted due to the occurrence of rapid radiations seem to have taken place (Robe et al. 2010).

Considering that *Mar* is a multiple copy sequence, the *Mar* phylogeny supports the view that the origin of this element might have occurred prior to divergence of these species and a burst of transposition have occurred at least in *D. willistoni*. Moreover, gene flow between *willistoni* subgroup species may also have contributed to the *Mar* evolutionary history. Besides the molecular support for introgression (Robe et al. 2010), there are evidences that some of the crosses between species within the *willistoni* subgroup produce fertile offspring (Cordeiro and Winge 1995).

Mar* copies from *D. willistoni

We identified 93 sequences (significant hits) of *Mar* in the *D. willistoni* genome. The exact number of copies is hard to determine because the genome contains some small and fragmented copies that are not rescued in the searches. Also, we cannot exclude the existence of duplicated scaffolds in the database, mainly the very short ones. From the sequences identified, 74 (79%) contain conserved TIRs of 11-bp, CAG(G/A)GGTAGGC, which are not perfect as found before (Holyoake and Kidwell 2003). Only one sequence presents perfect TIRs. The majority of copies (79%) present conserved TSDs of 8-bp, indicating recent mobilization of these sequences. The figure 3 shows the preferential insertion site (nTnTAnT/A) of *Mar*.

The analysis of the *Mar* copies distribution through the genome reveals that 32 copies are found in or near genes (Table S1 of supplementary material). This is consistent with previous reports on MITE association with genes (Wessler et al. 1995). Only one small region of *Mar* was found in a predicted coding sequence.

Exhaustive *in silico* searches were conducted in *D. willistoni* genome using the ends of *Mar* as queries in order to identify a possible transposase gene with some similarity to *Mar* or that share the same TIRs. However our searches did not reveal any candidate of full-length *Mar*.

Where did *Mar* come from and what is the transposase source for *Mar* mobilization?

Although it is not completely clear and may be distinct process can be involved in the origin of different MITE families, one hypothesis is that the MITEs are originated by deletion of autonomous copies. However, it is sometimes difficult to directly connect a given MITE family with an autonomous transposon present within the same genome. In many cases, sequence similarity between MITEs and the closest autonomous element is restricted to the TIRs (Feschotte and Prithan 2007).

Our searches in the *D. willistoni* genomes did not reveal any candidate of full-length *Mar* that could be the precursor of this MITE family. The autonomous element that gave rise to *Mar* may have been extinct in the genomes or in the analyzed strain. However, the conservation of TIRs and TSDs suggest recent mobilization of *Mar sequences*. Probably, *Mar* has been mobilized by the transposase of another element that present different TIRs sequence. Cross-mobilization is a highly associated process to amplification of MITE families (Yang et al. 2009). Examples are provided in rice, where the MITE *mPing* derived from the autonomous element *Ping* can be mobilized by a related autonomous element *Pong* (Jiang et al. 2003). Also, recently, Yang et al. (2009) showed cross-mobilization of MITEs from *Stowaway* family by *Osmar* transposase. In insects, within the *hAT* superfamily DNA transposons itself, cross-mobilization has been reported to the element *hobo* witch is able to mobilize the *hermes* transposon (Sundararajan et al. 1999).

A recent study (Ortiz and Loreto 2009) have characterized five different *hAT* families in *D. willistoni*, on which three are potentially active. We can suggest that some of these families of *hAT* element produce the transposase able to mobilize *Mar*. The table 2 summarizes the main characteristics of these *hAT* families according to Ortiz and Loreto (2009), and the figures 3 and 4 present a comparison among the TIRs and TSDs of *Mar* and these elements.

Other MITEs in *Drosophila*

It is known that MITEs are not very common elements in *Drosophila* genomes, or at least that they are not as abundant and diverse as in mosquitoes and plants. It’s important to note that the designation of MITE is not attributed to a common origin or a taxonomic level in TE classification. The designation of MITE is useful to describe this type of non-autonomous elements that share typical structural features: (1) short elements with no coding capacity; (2) can be present in a high number of copies. However, several MITE families have more modest number of copies (Quesneville et al. 2006; Grzebelus et al. 2009; Xu et al. 2010); (3) contain TIRs; (4) are often located in or near genes; (5) are AT-rich mainly the inner region. But an element should have all these features to be considered a MITE? Different authors can have different opinion about a non-autonomous element family.

POGON1 was the first element considered as a MITE in *Drosophila* (Feschotte et al. 2002), but other authors don’t agree with this designation (Kapitonov and Jurka 2002). There are around 25 copies of POGON1 in the euchromatin region of *D. melanogaster* genome and this element is a clear non-autonomous copy of POGO, formed by a big internal deletion and lack the inner region AT-rich that is typical of MITEs (Kapitonov and Jurka 2002). Could POGON1 be an intermediate step in a MITE family origin?

Other inconclusive instance is the element *DINE-I* (*dispersed repeat Drosophila interspersed element I*), which is the most abundant (>1000 copies) repetitive sequence in the *Drosophila* genome and it is found in all 12 *Drosophila* genomes. *DINE-I* was originally suggested to be relics of a family of ancient retroelements (Locke et al. 1999), and recently it was proposed that *DINE-I* is a family of MITE (Yang et al. 2006). However, some structure features of *DINE-I* support the proposal that these sequences are non-autonomous members of the *Helitron* order of TEs (Kapitonov and Jurka 2007). Elements in the order *Helitron* appear to replicate via a rolling-circle mechanism, with only one strand cut, and do not generate TSDs (Wicker et al. 2007). It's not clear yet if *DINE-I* is a MITE or a *Helitron* (Yang and Barbash 2008).

Vege is the other MITE found in *D. willistoni,* together with *Mar,* by Holyoake and Kidwell (2003). This element contains several characteristics of typical MITEs, however only few copies were found in the genomes (less than 10) by southern blot (Holyoake and Kidwell 2003). We confirm the presence of 3 copies of *Vege* in the *D. willistoni* genome that have imperfect 12-bp TIRs and conserved 8-bp TSDs (data not shown). These copies are not identical and have more than 5% of divergence with the described *Vege* sequence (GenBank: AF518730.1). Some other sequences (around 10) have similarities with *Vege* but do not have TIRs. The element *Vege* seems not very successfully spread throughout the genome compared to *Mar*.

The *hobo* element, from *hAT* superfamily, generates MITEs (*hobo*$^{\text{vahs}}$) in species from *melanogaster* subgroup (Ortiz and Loreto 2008). It was demonstrated that one of these MITEs is able to be mobilized by using the *transposase* of a *hobo* canonic (Torres et al. 2006), generating bursts of mutabilibity in *D. simulans* (Loreto et al. 1998). Recently, analyses of several autonomous *hAT* elements and their derivatives in the 12 *Drosophila* genomes suggest that several different families of MITEs were originated from different *hAT* elements (Ortiz et al. 2010).

The element *Mar* can be considered a *bona-fide* MITE in *Drosophila*. The source of transposase is unknown, but some candidates were presented in this work.

Mar evolutionary history

Mar sequences are present only in species from *Drosophila willistoni* subgroup. Based on the obtained data we can suggest that the origin of this MITE occurred after the separation of *willistoni* and *bocainensis* subgroups, but before the subgroup *willistoni* speciation that started about 5.7 Mya (Robe et al. 2010). Some related sequences may be present in the other species of the *D. willistoni* group, those from *bocainensis* subgroup.

We found several *Mar* copies in or near genes. Few of them have no conserved TIRs and/or TSDs and are more ancient insertions, may be present in the ancestor of *willistoni* subgroup. Nevertheless, most of gene-associated copies of *Mar* present conserved TIRs and TSDs indicating recent inserted copies that did not have enough time to accumulate mutations. This suggests that *Mar* can potentially be a powerful factor promoting intra and interspecies variability in *Drosophila*.

REFERENCES

Bureau TE and Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283-94.

Capy P, Bazin C, Higuet D and Langin T (1998) Dynamics and evolution of transposable elements. Landes Bioscience, Austin, Texas, 197 pp.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004) WebLogo: A sequence logo generator. Genome Research 14:1188-1190.

Cordeiro AR and Winge H (1995) Levels of evolutionary divergence of *Drosophila willistoni* sibling species. In: Levine L (ed) Genetics of natural populations: the continuing importance of Theodosius Dobzhansky. Columbia University Press, New York, pp 262-280.

Deprá M, Valente VL, Margis R and Loreto EL (2009) The *hobo* transposon and *hobo*-related elements are expressed as developmental genes in *Drosophila.* Gene 448:57-63.

Feschotte C and Pritham EJ (2007) DNA Transposons and the Evolution of Eukaryotic Genomes. Annu Rev Genet 41:331-368.

Feschotte C, Zhang X and Wessler SR (2002). Miniature inverted repeat transposable elements and their relationship to established DNA transposons. In: Craig NL, Craigie R, Gellert M and Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, DC, pp 1147-1158.

Gonzáles J and Petrov D (2009) Genetics. MITEs-the ultimate parasites. Science 325:1352-1353.

Grumbling G and Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Research 34:484-488.

Grzebelus D, Gładysz M, Macko-Podgórni A, Gambin T, Golis B, Rakoczy R and Gambin A (2009) Population dynamics of miniature inverted-repeat transposable elements (MITEs) in *Medicago truncatula*. Gene 448:214-220.

Holyoake A and Kidwell MG (2003) *Vege* and *Mar*: two novel *hAT* MITE families from *Drosophila willistoni*. Mol Biol Evol 20:163-167.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR and Wessler SR (2003) An active DNA transposon family in rice. Nature 421:163-167.

Kapitonov VV and Jurka J (2002) POGON1, a *bona fide* family of nonautonomous DNA. Repbase Reports 2:7.

Kapitonov VV and Jurka J (2007) *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521-529.

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626-1632.

Kidwell MG and Lisch DR (2001) Transposable elements, parasitic DNA, and genome evolution. Evolution 55:1-24.

Locke J, Howard LT, Aippersbach N, Podemski L and Hodgetts RB (1999) The characterization of *DINE-1*, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. Chromosoma 108:356-366.

Loreto EL, Zaha A, Nichols C, Pollock JA and Valente VLS (1998) Characterization of a hypermutable strain of *Drosophila simulans*. Cell Mol Life Sci 54:1283-1290.

Loreto ELS, Carareto CMA and Capy P (2008) Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100:545-554.

Ortiz MF and Loreto ELS (2008) The *hobo*-related elements in the *melanogaster* species group. Genet Res 90:243-252.

Ortiz MF and Loreto ELS (2009) Characterization of new *hAT* transposable elements in 12 *Drosophila* genomes. Genetica 135:67-75.

Ortiz MF, Lorenzatto KR, Corrêa BR and Loreto ELS (2010) *hAT* transposable elements and their derivatives: an analysis in the 12 *Drosophila* genomes. Genetica (in press).

Quesneville H, Nouaud D and Anxolabéhère D (2006) *P* elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. BMC Genomics 7:214.

Robe LJ, Cordeiro J, Loreto EL and Valente VL (2010) Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila williston*i subgroup (Diptera: Drosophilidae). Genetica (in press).

Robert CE (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Rubin E, Lithwick G and Levy AA (2001) Structure and evolution of the *hAT* transposon superfamily. Genetics 158:949-957.

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Sassi AK, Herédia FO, Loreto ELS, Valente VLS and Rohde C (2005) Transposable elements *P* and *gypsy* in natural populations of *Drosophila willistoni*. Genet Mol Biol 28:734-739.

Sundararajan P, Atkinson PW and O'Brochta DA (1999) Transposable element interactions in insects: crossmobilization of *hobo* and *Hermes*. Insect Mol Biol 8:359-368.

Staden R (1996) The Staden sequence analyses package. Mol Biotechnol 5:233-241.

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599.

Thornburg BG, Gotea V and Makayowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104-110.

Torres FP, Fonte LFM, Valente VLS and Loreto ELS (2006) Mobilization of a *hobo*-related sequence in the genome of *Drosophila simulans*. Genetica 123:101-110.

Wessler SR, Bureau TE and White SE (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814-821.

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

Xu J, Wang M, Zhang X, Tang F, Pan G and Zhou Z (2010) Identification of NbME MITE families: potential molecular markers in the microsporidia *Nosema bombycis*. J Invertebr Pathol 103:48-52.

Yang G, Nagel DH, Feschotte C, Hancock CN and Wessler SR (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. Science 325:1391-1394.

Yang HP, Hung TL, You TL, Yang TH (2006) Genome-wide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. Genetics 173:189-196.

Yang HP and Barbash DA (2008) Abundant and species-specific *DINE-1* transposable elements in 12 *Drosophila* genomes. Genome Biol 9:R39.

TABLES

Table 1: Drosophilidae species investigated in this work with their taxonomic placement and their respective PCR and dot blot results.

Genus	Subgenus	Group	Species	PCR	Dot blot
Drosophila	*Drosophila*	*guarani*	*D. ornatifrons*	-	-
			D. subbadia	-	-
			D. guaru	-	-
		guaramuru	*D. griseolineata*	-	-
			D. maculifrons	-	-
		tripunctata	*D. nappae*	-	-
			D. paraguayensis	-	?
			D. crocina	-	-
			D. paramediostriata	-	-
			D. tripunctata	-	-
			D. mediodiffusa	-	?
			D. mediopictoides	-	-
		cardini	*D. cardinoides*	-	?
			D. neocardini	-	-
			D. polymorpha	-	-
			D. procardinoides	-	?
			D. arawakana	-	?
		pallidipennis	*D. pallidipennis*	-	?
		calloptera	*D. ornatipennis*	-	-
		immigrans	*D. immigrans*	-	-
		funebris	*D. funebris*	-	-
		mesophragmatica	*D. gasici*	-	-
			D. brncici	-	?
			D. gaucha	-	-
			D. pavani	-	?
		repleta	*D. hydei*	-	-
			D. mercatorum	-	-
			D. mojavensis	-	-
			D. buzzati	-	?
		canalinea	*D. canalinea*	-	-
		flavopilosa	*D. cestri*	-	?
			D. incompta	-	-
		virilis	*D. virilis*	-	-
		robusta	*D. robusta*	-	-
	Sophophora	*melanogaster*	*D. melanogaster*	-	-
			D. simulans	-	-
			D. sechellia	-	?
			D. mauritiana	-	-
			D. teissieri	-	-
			D. santomea	-	-
			D. erecta	-	-
			D. yakuba	-	-
			D. kikkawai	-	-
			D. ananassae	-	-
			D. malerkotliana	-	-
			D. orena	-	-
		obscura	*D. pseudoobscura*	-	-
		saltans	*D. prosaltans*	-	-
			D. saltans	-	-
			D. neoelliptica	-	-
			D. sturtevanti	-	-
		willistoni	*D. sucinea*	-	w
			D. nebulosa	-	-
			D. capricorni	-	w
			D. fumipennis	-	w

		D. willistoni *	+	+
		D. paulistorum *	+	+
		D. insularis	+	+
		D. tropicalis	+	+
		D. equinoxialis	+	+
	Dorsilopha	*D. busckii*	-	-
Zaprionus		*Z. indianus*	-	-
		Z. tuberculatus	-	-
Scaptodrosophila		*S. latifasciaeformis*	-	-
		S. lebanonensis	-	-

* More than one strain was used for these species. *D. willistoni* strains: ww, 17A2 and WIP4. *D. paulistorum* strains: Ori (semispecies Orinocan), Andi and PR (semispecies Andean-Brazilian).
(-) no amplification or hybridization signal; (+) positive amplification or hybridization; (w) weak hybridization signal; (?) not tested.

Table 2: Characteristics of five families of *hAT* transposons in *D. willistoni*

	Family	Size (bp)	TIRs (bp)	TSDs (bp)	Copy number (full-lenght)
Howilli1	*herves*	2,816 (pa)	11	8	2
Howilli2	*hobo*	2,847 (pa)	13	8	2
Howilli3	*homo*	2,559 (pa)	12	8	2
Howilli4	Unclustered	2,631 (pna)	11	8	3
Howilli5	*hobo*	2,401 (pna)	11	8	3

pa – potentially active; pna – potentially non active

FIGURES

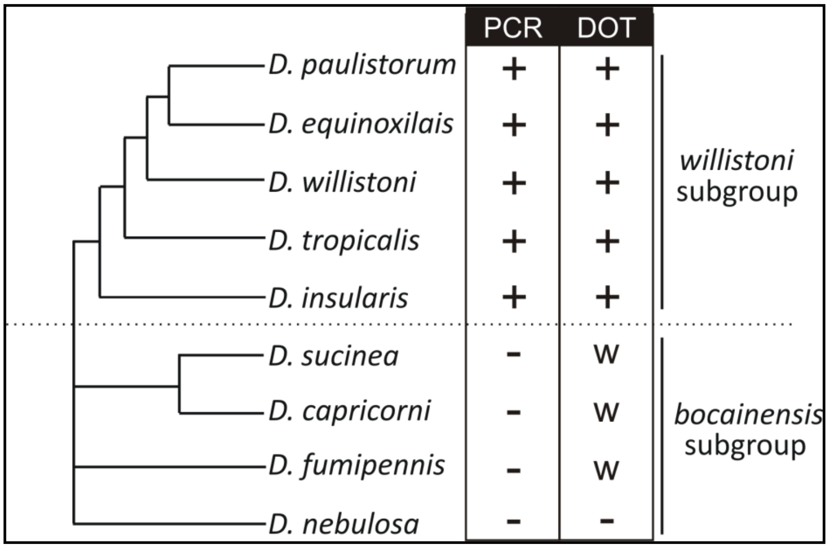
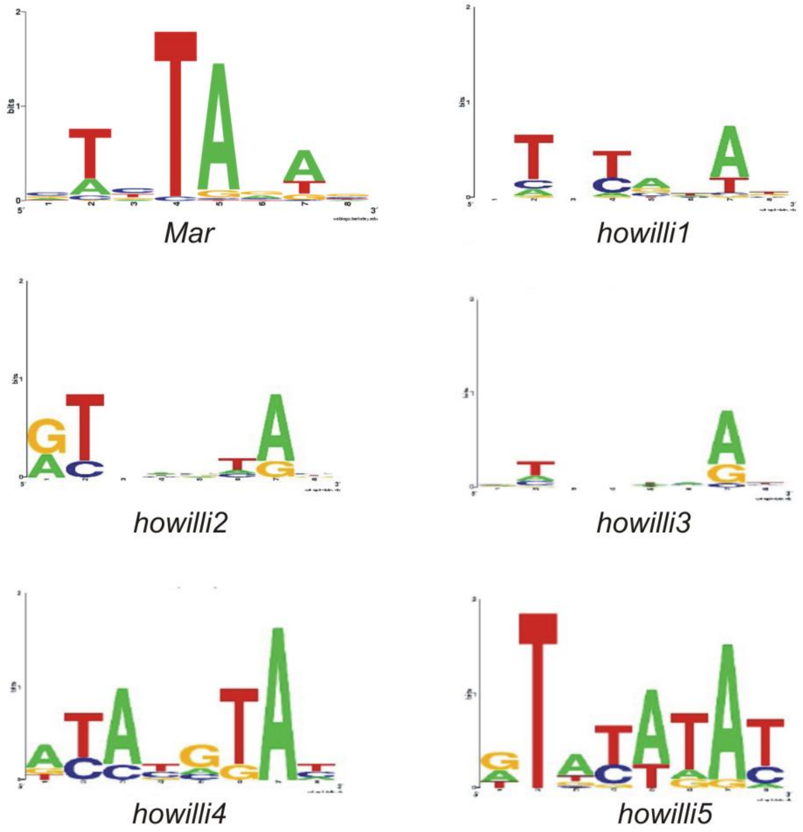


Figure 1: Evolutionary relationships of the *willistoni* group (based on Robe et al. 2010) and the results observed in the PCR and dot blot. + positive result; - negative result; w - weak signal.



Mar

howilli1

howilli2

howilli3

howilli4

howilli5

Figure 2: Sequence logo analysis for the 8 bp TSDs observed for *Mar* and for the *hAT* elements found in *D. willistoni* (Ortiz et al. 2010). The y-axis shows 2 bit of information and the x-axis represents the nucleotide position in the TSDs.

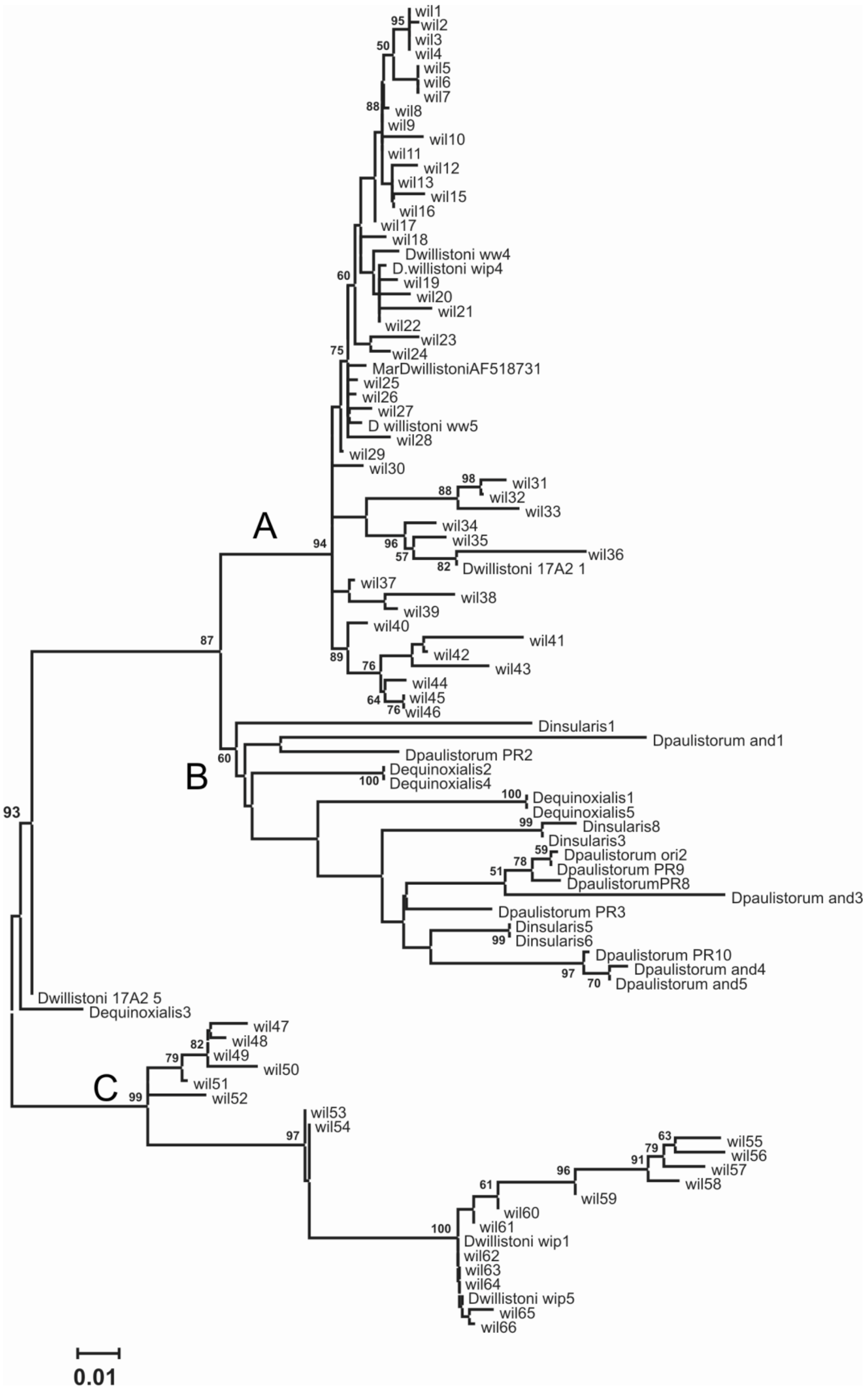
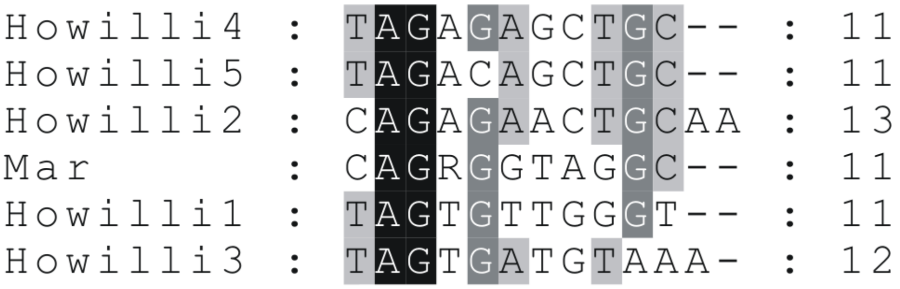


Figure 3: Neighbor-Joining tree of *Mar*. Bootstrap values are shown at nodes and those smaller than 50 were omitted. Sequences from *D. willistoni* genome are named “wil” followed by a number. Sequences originated by cloning are named as: species name_ strain_ number of the clone.



```
Howilli4 : TAGAGAGCTGC-- : 11
Howilli5 : TAGACAGCTGC-- : 11
Howilli2 : CAGAGAACTGCAA : 13
Mar      : CAGRGGTAGGC-- : 11
Howilli1 : TAGTGTTGGGT-- : 11
Howilli3 : TAGTGATGTAAA- : 12
```

Figure 4: Alignment of TIR sequences from *Mar* and *hAT* elements from *D. willistoni*. The major similarity with *Mar* TIRs is presented *by howilli1* with 6 matches from 11.

SUPPLEMENTARY MATERIAL

Table S1: Putative genes that contain or are near *Mar* copies.

Gene	Score	e-value	Ortholog	Domains or function
Dwil\GK23646	504	e-142	-	RT
Dwil\GK16085	496	e-139	-	RH, LIM
Dwil\GK10438	496	e-139	-	IG
Dwil\GK21996	480	e-135	+	ecdysone receptor
Dwil\GK16111	470	e-132	+	Utp11, Ufd2P
Dwil\GK12298	456	e-127	+	glutamate receptor
Dwil\GK22081 (mastermind)	446	e-124	+	Transcription coactivator
Dwil\GK18788	446	e-124	+	ubiquitin thiolesterase
Dwil\GK18750 (semaphorin)	446	e-124	+	receptor activity
Dwil\GK17026	446	e-124	-	thioredoxin domain
Dwil\GK23613	203	e-111	-	flgl
Dwil\GK20297	359	2e-098	-	Piwi-like domain
Dwil\GK18552	353	1e-096	-	Transcription repressor
Dwil\GK10543	307	6e-083	+	microtubule binding
Dwil\GK10422 (chico)	289	1e-077	+	insulin-like growth factor receptor binding
Dwil\GK16644	252	3e-066	-	Importin domain
Dwil\GK23093 (dystroglycan)	186	2e-046	+	Protein binding
Dwil\GK15940	519	e-146	+	gustatory receptor
Dwil\tRNA:GK26087	480	e-134	-	-
Dwil\GK10602	478	e-134	-	Homologue to exon eIF-5A
Dwil\GK21043	468	e-131	-	Gypsy2-I_Dmoj
Dwil\GK23197	446	e-124	-	unknow
Dwil\GK14562	446	e-124	-	DUF3743
Dwil\GK10515	446	e-124	-	unknow
Dwil\GK14495	438	e-122	-	unknow
Dwil\GK10612	385	e-105	-	unknow
Dwil\GK18774	379	e-104	-	unknow
Dwil\GK10531 (omega)	353	3e-096	+	dipeptidyl-peptidase activity
Dwil\GK25289	313	3e-084	+	branched-chain-amino-acid transaminase activity
Dwil\GK16590	297	2e-079	-	unknow
Dwil\tRNA:GK26340	208	1e-052	-	-
Dwil\GK17166	240	4e-062	-	AdoMet_MTase
Dwil\GK25464	230	3e-059	-	unknow
Dwil\tRNA:GK26349	208	1e-052	-	-
Dwil\GK20176	178	1e-043	-	unknow
Dwil\GK12334	170	3e-041	+	Unknow

Gene function was checked in the flybase gene database or the conserved domains were examined using Pfam database (Finn et al. 2010. Nucleic Acids Research; Database Issue 38:D211-222). In the ortholog column, (+) means presence of otholog genes in other species and (-) means absence of otholog genes in other species.

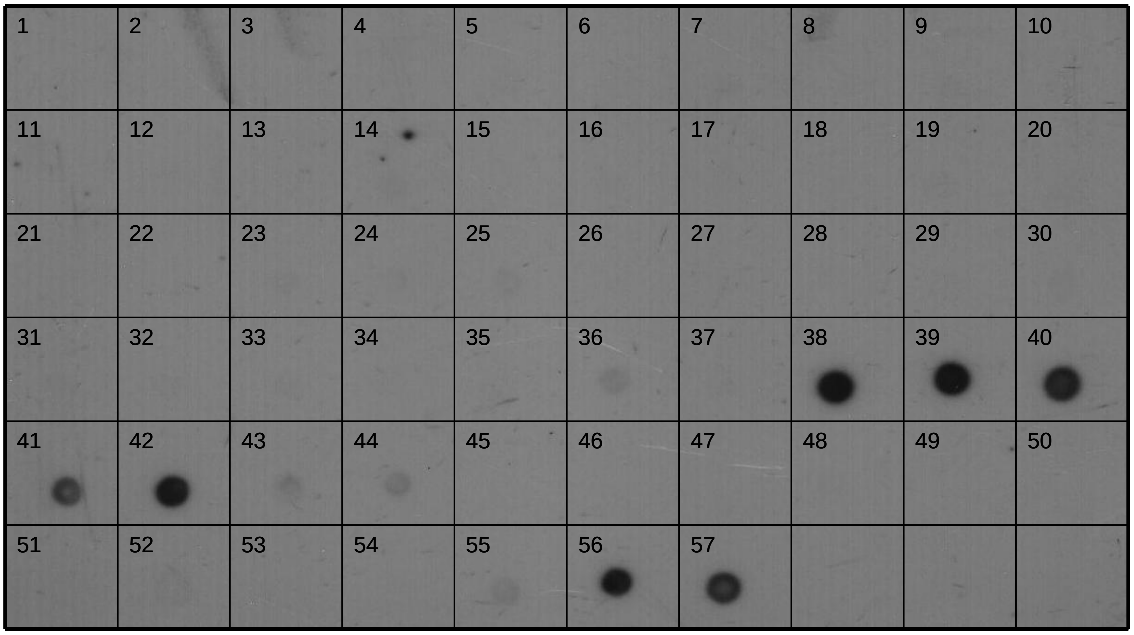


Figure S1: Dot blot screening for presence of *Mar*. The species tested are the followed: *1 – D. ornatifrons; 2 – D. subbadia; 3 – D. guaru; 4 – D. griseolineata; 5 – D. nappae; 6 – D. paramediostriata; 7 – D. tripunctata; 8 – D. medipictoides; 9 – D. neocardini; 10 – D. polymorpha; 11 – D. ornatipennis; 12 – D. immigrans; 13 – D. funebris; 14 – D. gasici; 15 – D. gaucha; 16 – D. mercatorum; 17 – D. mojavensis; 18 – D. incompta; 19 – D. virilis; 20 – D. robusta; 21 – D. melanogaster; 22 – D. simulans; 23 – D. mauritiana; 24 – D. teissieri; 25 – D. santomea; 26 – D. erecta; 27 – D. yakuba; 28 – D. kikkawai; 29 – D. ananassae; 29 – D. malerkotliana; 30 – D. orena; 31 – D. pseudoobscura; 32 – D. prosaltans; 33 – D. saltans; 34 – D. neoelliptica; 35 – D. sturtevanti; 36 – D. sucinea; 37 – D. nebulosa; 38 - D. willistoni (ww); 39 – D. paulistorum Orinocan; 40 – D. insularis; 41 - D. tropicalis; 42 - D. equinoxialis; 43 - D. capricorni; 44 – D. fumipennis; 45 – D. busckii; 46 – Z. indianus; 47 – Z. tuberculatus; 48 – S. latifasciaeformis; 49 – S. lebanonensis; 50 – without DNA; 51 – D. maculifrons; 52 – D. crocina; 53 – D. hydei; 54 – D. canalinea; 55 – D. orena; 56 – D. willistoni (Wip4); 57 – D. willistoni (17A2)*

Capítulo 5

Evolutionary history of the *Kanga* retroviral lineage in *Drosophila**

Adriana Ludwig[1], Elgion L. S. Loreto[1,2], Harmit Singh Malik[3,4]

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil.

[2] Departamento de Biologia, Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brazil. Email: elgion.loreto@pq.cnpq.br

[3] Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle Washington, United States of America. Email: hsmalik@fhcrc.org

[4] Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle Washington, United States of America.

* Este trabalho foi realizado pela aluna Adriana Ludwig com bolsa de doutorado sanduíche pelo CNPq.

Manuscrito em preparação para ser enviado à revista Gene.

ABSTRACT

Mobile elements can have dramatically deleterious consequences on host genomes, but occasionally their genes are usurped by host genomes because they are selectively advantageous. In *Drosophila*, the *Iris* gene is a domesticated retroviral *env* gene that was originated from a previously unknown lineage of *Drosophila* endogenous retroviruses, *Kanga*. Aiming to contribute to knowledge of this domesticated process we performed evolutionary analyses of the progenitor retrovirus. We have performed *in silico* searches and PCR screening for *Kanga* homologous sequences in diverse *Drosophila* species. The absence of *Iris* gene in *D. willistoni* is in accordance with its origin in the ancestor of *melanogaster* and *obscura* groups around 25 million years ago. We found three ancient lineages of *Kanga-like* retroviruses that, in some *Drosophila* species, still appear to have active copies, which underwent recent retrotransposition, as can be suggested by structure and LTR conservation. *Kanga2* lineage, which is the most closely related to *Iris* gene, has a very restricted distribution, being present only in some closely related species to *D. ananassae*. Putative *Kanga2* envelope protein has conserved the features conferring fusogenic property, which is required for infectivity. *Kanga-like* endogenous retroviruses have been transmitted vertically and have a long term coevolution with *Drosophila* genomes, being an important model of a mutualistic relationship between transposable element and their host.

Keywords: *Iris*, *Kanga*, endogenous retroviruses, retrotransposon, envelope, domestication

INTRODUCTION

LTR-retroelements are among the most abundant constituents of eukaryotic genomes and have had important role in the course of evolution, shaping the size, structure and expression of their host genomes. Based on organization of domains and reverse transcriptase phylogeny the LTR-retroelements can be divided into 4 groups: *Ty1-copia*, *Ty3-gypsy*, *Bel-Pao* and vertebrate retrovirus (Eickbush and Jamburuthugoda 2008; Llorens et al. 2009). Their integrated proviral form consist of two long terminal repeats (LTR) flanking an internal region that contains two or three genes, *gag, pol* and *env* (Figure1).

The *gag* gene encodes structural proteins that form the virus-like particle, inside which reverse transcription takes place. The *pol* gene encodes enzymes, including a protease that cleaves the Pol polyprotein, a reverse transcriptase that copies the retroelements RNA into cDNA and an integrase that integrates the cDNA into the genome. The *gag* and *pol* genes are believed to be necessary and sufficient for transposition and intracellular life cycle. However, retrovirus and a number of retrotransposons from the other groups have an envelope-like gene that is responsible for the infectious ability of retroviruses (Havecker et al. 2004). Retroviruses enter the host cell through fusion of virus and cell membranes which requires a proteolytic cleavage to separate the envelope (Env) protein into the SU (receptor-binding component) and TM (brings membranes into close apposition and causes fusion) proteins. Just downstream of the furin cleavage site in the Env protein is a hydrophobic fusion peptide that is also required for membrane fusion (Coffin et al. 1997).

Envelope genes have been acquired independently on several occasions during the evolution of LTR-retroelements (Malik et al. 2000). Two of these instances led to *gypsy-like* (*Ty3-gypsy*) and *roo-like* (*Bel-Pao*) retrotransposons in *Drosophila*, which have both separately, acquired homologous *env* genes from baculoviruses, insect pathogenic double-stranded DNA viruses (Malik et al. 2000; Malik and Henikoff 2005). Whereas vertebrate retroviruses are predominantly transmitted horizontally by cell-to-cell infection, these *Drosophila* retrotransposons, also known as endogenous retroviruses, are mainly transmitted vertically from mother to offspring as integrated copies in the host cell genome (Chalvet et al. 1999; Huszar and Imler 2008). However, studies of *gypsy-like* retrotransposons have shown that during the evolution and diversification of these elements, the *env* gene has been conserved indicating that infectious capacity have been important to the survival and expansion of these retroviruses in the genomes (Leblanc et al. 2000; Llorens et al. 2008; Ludwig et al. 2008; Mejlumian et al. 2002).

Although transposition activity can cause a broad spectrum of disadvantageous mutations, the genomes have coevolved with their transposable elements (TEs), devising strategies to prevent them while co-opting function from their presence (Volff 2006). The “molecular domestication”, when a TE coding sequence gives rise to a functional host gene, is probably the most striking beneficial contribution of TEs. Malik and Henikoff (2005) identified a relatively recent domestication event of a retroviral *env* gene in *Drosophila.* This gene, named *Iris*, is highly conserved and it is expressed in both adult male and female. Other important characteristics of this gene are the lack of the cleavage site and the action of strong positive selection on it. The function of *Iris* is unknown at this stage. The authors suggested that *Iris* have been recruited by the host genome to combat retrovirus infection.

Iris was domesticated from a novel family of *Bel-Pao* retrotransposons, called *Kanga*, which is most closely related to the *roo* retrotransposon (Malik and Henikoff 2005). In this work, we show that *Kanga-like* endogenous retroviruses are ancient components of the *Drosophila* genomes, although they still have some copies that are potentially active.

MATERIAL AND METHODS

In silico searches and sequence structure analyses

Searches for sequences homologous to *Kanga env* gene and *Iris* gene were performed in the 12 available *Drosophila* genomes: *D. melanogaster, D. sechellia, D. simulans, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis* and *D. grimshawi*. We used the BLAST tool available on the Flybase (http://flybase.bio.indiana.edu/blast/; Grumbling and Strelets 2006). TblastN were performed using the *kanga2 env* amino acid sequence from *D. ananassae* (Malik and Henikoff 2005) as query. All sequences found with e-value $<10^{-4}$ and more than 70% of the query length were examined for the presence of LTRs, primer binding sites (PBS) and target site duplication (TSDs) using LTR finder (Xu and Wang 2007) and visual inspection. The ORF finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) was used to determine the ORF structure. Kyte-Doolittle hydropathy plots (Kyte and Doolittle 1982) of predicted Env protein were used to identify the signal peptide, transmembrane domain and fusion peptide. SMART tool (Letunic et al. 2009; http://smart.embl.de/) was also used to identify these domains.

Phylogenetic analyses

The amino acid sequences were deduced by GeneDoc 2.6.001 program (Nicholas and Nicholas, 1997). Amino acid sequences were aligned using the program MUSCLE (Edgar 2004). The amino acid phylogeny was obtained by Neighbor-joining in the Mega4 software (Tamura et al. 2007), using default parameters and 1000 replication bootstrap.

Nucleotide sequences of *Kanga2 env* gene obtained by *in silico* searches and by cloning (see below) were aligned using the program MUSCLE (Edgar 2004). The nucleotide phylogeny was obtained by Bayesian analysis implemented in the MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003) using the model GTR+G indicated by MrModeltest 2.3 (Nylander 2004).

Horizontal transfer test

In order to test horizontal transfer (HT) hypothesis, we compared the synonymous substitutions (dS) rate between *Kanga* and the nuclear gene *alpha methyldopa* (*amd*). The dS values offer a measurement of neutral evolution in the absence of strong codon usage bias. Therefore, the same proportion of dS should be expected for TEs and host genes when two species are compared. If a TE has been transmitted from one species to another via a recent HT, the dS value for the TE should be significantly lower than the dS for the host genes. Nevertheless, differences in the magnitude of dS among genes have been found to be negatively correlated to the intensity of natural selection on synonymous codon usage (Shields et al. 1988; Sharp and Li 1989). Thus, different authors have been used the comparisons of dS to infer HT, but considering also the level of codon bias.

We used the *amd* gene for the dS comparisons since it was observed that this gene is very useful for that, as it does not present a very high codon usage bias (Ludwig et al. 2008; Vidal et al. 2009). If this were the case, very low dS values would be observed, which would in turn lead to an underestimation of HT occurrence. Apart from this, retrotransposons generally present a lower degree of codon usage bias than the *amd* gene, which results in higher dS values and avoids overestimation of HT occurrence.

Codon alignments for *Kanga1A*, *Kanga1B* and *amd* were used to estimate the dS values using the Nei and Gojobori (1986) method assisted by the MEGA 4 software (Tamura et al. 2007). The effective number of codons (Nc; Wright 1990) and the codon bias index (CBI; Morton 1993) were calculated by DNASP 4.0 software (Rozas et al. 2003).

Insertion time estimative

Time of insertion was calculated using the formula $T = k/2r$, where T is the time of divergence, k is the divergence between the LTRs, and r is the substitution rate (Graur and Li 2000). The substitution rate used was 0.016 substitutions per site per million years as calculated for *Drosophila* genes with low codon usage bias (Sharp and Li 1989).

PCR screening, cloning and sequencing of *Kanga2*

The list of 81 *Drosophila* species screened is shown in the table S1 of supplementary material. DNA from adult flies was prepared using the DNAeasy Blood and Tissue Kit (Qiagen). PCR approach was used to amplify the entire *env* gene from *Kanga2* using the following primers: Fk2F 5’-GCGGTGTTACCAATCAACG-3’ and Fk2R 5’-GATAGTGGTCACTCTTCGATGG-3’. PCRs were carried out with the Supermix High Fidelity (Invitrogen) in several different amplification conditions. PCR products were cloned using the TOPO TA cloning kit (Invitrogen), followed by sequencing of clones using vector primers and internal primers for *kanga2*. Assembly of clone’s sequences was performed by the electropherogram analyses using the GAP 4 software of the Staden Package (Staden 1996).

Southern blot

DNA samples were digested with *Eco*RI, electrophoresed in 1.2% agarose gel and transferred to a nylon filter. The probe region is shown in the figure 2A and was obtained by PCR of a *Kanga2* clone and labeled with Gene Images AlkPhos Direct labeling system (GE Heathcare). Hybridization and detection were performed in according to the AlkPhos Direct labeling and detection Kit. Signal detection was performed using CDP-star reagent followed by exposure to autoradiographic film.

RESULTS AND DISCUSSION

Searches for *Kanga* homologous sequences

In order to characterize the new endogenous retrovirus lineage, *Kanga,* and to better understand the origin of *Iris* gene we initiate searches in the 12 *Drosophila* genomes. The results of *in silico* searches are summarized in the table 1. The *Iris* gene was found in syntenic location only in species from the sister groups, *melanogaster* and *obscura*. The other species have no true orthologous, even in other genomic location. These are in accordance with the origin of this gene in the ancestor of *melanogaster* and *obscura* groups around 25 million years ago (Malik and Henikoff 2005). The *melanogaster* subgroup species have no *kanga-like* sequence. The best hits sequences, besides the *Iris* gene, match to the related endogenous retrovirus, *roo*. *D. virilis* genome also does not contain any *kanga-like* sequence. All other species have some sequence homologous to *Kanga* retroviruses.

Structural features and evolution of *Kanga-like* retroviruses

More than 100 sequences were analyzed and the ones that have complete *env* gene were used to infer the evolutionary relationships among *Iris* and *Kanga-like env* gene (Figure 3). The phylogeny is divided into 4 groups: The *Iris* gene clade and 3 *Kanga-like* clades*, Kanga1A*, *Kanga1B* and *Kanga2*. We have performed structural analysis of these endogenous retroviruses (Table 2).

The groups *Kanga1*A and *Kanga1B* are very closely related, but were divided into two distinct groups considering the distribution of these sequences and phylogeny. *Kanga1A* contains sequences from *D. pseudoobscura, D. persimilis, D. grimshawi* and *D. ananassae.* Most of sequences from *D. pseudooscura* are arranged as *in tanden* repeats in a 130 Kb scaffold. The presence of LTRs was not detected, but some copies are composed by one ORF containing all encoding regions, *gag*, *pol* and *env*. Majority of the copies from *D. persimilis* are located in incomplete or short scaffolds making hard to conclude about the structure and conservation of these copies, but apparently, the copies have degenerated ORFs. *D. grimshawi* presents one full-length copy which contains LTRs 96.5% similar and conserved TSDs, but the ORFs are degenerated. *D. ananassae* presents some full-length copies and a consensus sequence presents one unique ORF encoding for all genes although the individual copies have internal stop codons.

Kanga1B retrovirus is present in *D. ananassae, D. willistoni, D. mojavensis, D. persimilis* and *D. pseudoobscura.* One of the sequences found in *D. ananassae* is potentially active since contains identical LTRs, conserved TSDs and one large ORF containing all domains. The *Kanga1B* sequences from *D. willistoni* are also potentially actives. They have conserved LTRs and TSDs, but the scaffolds are incomplete presenting missing data regions in the middle of the sequences; however, a consensus sequence shows a conserved ORF encoding for the three genes. Sequences from *D. mojavensis* and *D. pseudoobscura* and *D. persimilis* are apparently inactive, but since many sequences are found in short scaffolds or in the extremity of those, it is hard to conclude about the structure of these sequences. One full-length sequence is found in *D. persimilis* and *D. mojavensis*, however the ORFs are apparently degenerated.

The *Kanga2* was found only in *D. ananassae* and it is the most related clade to the *Iris* gene. Some complete sequences and potentially active can be found. This retrovirus presents different expression strategy, in which the *gag* and *pol* genes are fused together in the same frame, while the *env* gene has a single ORF.

A representative or a consensus sequences for all *Kanga-like* lineages were used to determine some structure features in each species (Table2). All *Kanga-like* lineages have TSDs of 5-bp that is a common size for *Bel-Pao* elements (Minervini et al. 2009). Total and LTR lengths differ in each species.

Located immediately downstream of the LTR 5' is an 18-bp long primer binding site (PBS), complementary to the 3'-sequence of a host tRNA, which is used as a primer for initiation of reverse transcription (Coffin et al. 1997). Almost all characterized copies have a PBS site complementary to tRNA^{Trp} or tRNA^{Arg}.

We also analyzed the deduced *Kanga-like* Env proteins. One feature that is almost universally conserved among retroviral Env proteins is a furin cleavage site (R-X-X-R; where R is Arginine and X is any amino acid; Krysan et al. 1999) that separates SU from TM (Coffin et al. 1997). The putative furin cleavage site for the *Kanga-like* retroviruses is also indicated in the table 2.

The relative integration time or age of an LTR-retroelement can be estimated from the level of divergence existing between the LTRs of a copy, since the LTRs are identical at the DNA sequence level on integration (Bowen and McDonald 2001). As can be seen in the table 2, most of the *kanga-like* insertions are very new (0 to 1.7 million years ago), suggesting recently activity of these sequences. However, this divergence method to estimate the insertion time should be applied with special caution because at least some LTRs undergo gene conversion (Kijima and Innan 2010). Moreover, in this case, the age of the retroviruses in the species is not the same as the integration time, because we found several degenerated copies (without LTRs) of all *kanga* lineages. It suggests that these retroviruses may be ancient components of *Drosophila* genomes.

To better understand the evolutionary relationship of *Kanga-like* retroviruses, we compare their phylogeny (Figure 3) with the host species phylogenetic tree (Figure 4). *Kanga-like* sequences have a patchy distribution in the *Drosophila* phylogeny. Three hypotheses could explain this pattern: (1) some species have lost *Kanga* sequences during the evolution; (2) polymorphic copies were present in the ancestor and independently assorted during speciation; (3) some copies could have been horizontally transmitted between species.

To test the HT hypothesis we compared the dS values found for *Kanga* and the host gene *amd*. As dS values offer a measurement of neutral evolution in the absence of a strong codon usage bias, consequently, the same proportion of TEs and host genes dS should be expected (see Material and Methods). Tables 3 and 4 show the dS comparison for *Kanga1A* and *Kanga1B*, respectively. As expected by vertical transmission, the dS values from *Kanga1A* and *Kanga1B* are very similar to those from the *amd* gene. In addition, *Kanga1A* and *Kanga1B* show a similar or slightly lower level of codon usage bias than *amd* (Table 5), which avoid differences in the degree of dS between genes caused by selection in the synonymous codons.

Characterization and evolution of *Kanga2*

Kanga2 is the most related clade to the *Iris* gene (Figure 3) and we focused some analysis in this lineage. *Kanga2* has 9877-bp long with 1015-bp LTRs (Figure 2A). We found a PBS sequence complementary to tRNA^{Trp}. The central part comprises three genes corresponding to the *gag, pol* and *env*. In the *in silico* searches, *kanga2* was restricted to *D. ananassae*, which could indicate that this lineage has a more recent origin than *Kanga1* lineage. Most of 24 sequences of *kanga2* are degenerated due the absence of LTRs, presence of unexpected stop codons, frameshifts, insertions or deletions. However, 6 copies are full length and 3 are potentially active, with conserved encoding regions.

The *env* gene potentially encodes a functional envelope protein (Figure 2B), that presents all features to have fusion and infectivity properties: (1) a signal peptide that is proteolytically removed by the action of a cellular protease within the endoplasmic reticulum; (2) a putative furin-like cleavage site present between the Surface and Transmembrane domains. SU protein serves to bind the virus to the host cell, and the TM protein serves both to anchor the entire viral glycoprotein complex on the virion surface and to mediate the fusion of the virion with the host-cell membrane during entry; (3) a hydrophobic fusion peptide that is exposed after the cleavage of SU and TM; (4) potential N-glycosylation sites, that can vary in number and distribution between different retroviruses; (5) several cysteine residues, that potentially participate in a disulfide bond either within the same molecule or across different molecules.

In order to investigate the evolution of *Kanga2* in *Drosophila*, we looked for its presence in other 81Drosophilidae species using PCR. *Kanga2* was found only in *D. ananassae* (as evidenced *in silico*), and in the sister species, *D. pallidosa*, *D. pallidosa-like* and *D. pallidosa-like Wau*. These species are very closely related to *D. ananassae*, but are partially reproductively isolated and have distinct chromosome arrangements (Matsuda et al. 2009).

In the *kanga2* phylogeny (Figure 5), there is very little clustering by species, as the sequences from *D. ananassae* and *D. pallidosa* are scattered in the phylogeny. This indicates that the diversification of *Kanga2* might have occurred prior to divergence of these species. Also, gene flow between these species may also have contributed to the *Kanga2* evolution, since low pre-mating isolation is observed between *D. ananassae* males and *D. pallidosa* females (Matsuda et al. 2009). On the other hand, *D. pallidosa-like Wau* shows strong pre-mating isolation from the other *ananassae* complex taxa (Matsuda et al. 2009). In accordance with low gene flow involving *D. pallidosa-like Wau*, its *kanga2* sequences form a very well separated clade in the *nik* phylogeny.

We also performed Southern blot (Figure 6) using different populations of *D. ananassae* and *D. pallidosa*. The enzyme EcoRI was chosen after analysis of the sequence of several clones. This enzyme cleaves the *env* gene in a single site and has no cleavage site in the 3 'LTR. The cleavage site and the probe regions can be seen in Figure 2A. Excluding the possibility of polymorphism of the cleavage site, each band should represent one copy. Multiple copies were observed in the *D. ananassae* and *D. pallidosa* genomes with most bands shared between strains, indicating that they are copies present in the ancestor of both species.

Concluding Remarks

Endogenous retroviruses are reminiscences of ancient retroviral infections of the host germline transmitted vertically from generation to generation. They are widespread in many animal species and provide a powerful source of genomic variation and have contributed to the generation of genomic novelties (Mourier and Willerslev 2009; Varela et al. 2009).

In this work, we present the evolutionary study of the *Drosophila* endogenous retroviruses *Kanga* that belongs to the Bel-Pao group of LTR-retroelements.

We found three, ancient lineages of *Kanga-like* elements that still appear to have at least some active copies in some *Drosophila* species. Our data suggest that *Kanga* retrovirus may have originated in the ancestor of *Drosophila* and *Drosophila* subgenera, more than 40 million years ago. The split of *Kanga1A* and *Kanga1B* probably have occurred before the separation of these subgenera. One may suggest that these endogenous retroviruses originally had infectious capacity afforded by the *env* gene, although the main mode of *Kanga-like* transmission seems to have been vertically. However, we cannot discard horizontal transfer events involving species not investigated in this work. Although *kanga1* lineages are very old, some species show evidence of recent mobilization and structural characteristics of functional retroviruses.

Events of stochastic loss can explain the absence of *kanga-like* sequences in the *melanogaster* subgroup species and *D. virilis,* as well as the absence of

Kanga1A in *D. willistoni* and *D. mojavensis* and *Kanga1B* in *D. grimshawi*. Different assortment of polymorphic copies during speciation process also could explain the distribution of *Kanga-like* sequences. The closely relationship between *kanga2* and *Iris* gene suggest that *Kanga2* could be present in the ancestor of *melanogaster* and *obscura* groups, where the *Iris* gene was domesticated. However, *Kanga2* has a very limited distribution and should have been lost in all *Drosophila* species except *ananassae* complex species. It implies several independent and some recent lost events. A more plausible explanation is that *Kanga2* was lost anciently in the *Drosophila* evolution and may have invaded the genome of the *ananassae* complex species more recently coming from a unidentified source that have maintained *Kanga2* since that.

An important fact in the evolution of endogenous retroviruses *Kanga-like* was the domestication of the *env* gene, about 25 million years ago in the ancestor of the *melanogaster* and *obscura* groups. Possibly by acquiring an important role in the organism, this *env* gene called *Iris*, has been kept since then in the genomes whereas the other ORFs were being degenerated. The function of *Iris* genes remains to be characterized. Several examples of domestication of retroviral genes are described in vertebrates and, in most instances, the donor is an extinct retrovirus (Varela et al. 2009). However, it is different in the case of *Drosophila Iris* gene, where *Kanga* retroviruses seem to be still active in at least some *Drosophila* species, co-existing with their derived *Iris* gene.

Our results suggest that endogenous retroviruses can be maintained for long evolutionary periods in *Drosophila*, possibly by a balance among positive, negative, and neutral selective influences. This long term co-evolution of endogenous retroviruses *Kanga-like* and the *Drosophila* genomes is an important example of a mutualistic relationship between TEs and their host.

REFERENCES

Bowen NJ and McDonald JF (2001) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11:1527-1540.

Chalvet F, Teysset L, Terzian C, Prud'homme N, Santamaria P, Bucheton A and Pélisson A (1999) Proviral amplification of the *Gypsy* endogenous retrovirus of *Drosophila melanogaster* involves *env*-independent invasion of the female germline. EMBO J 18:2659-2669.

Coffin JM, Hughes SH and Varmus HE (1997) Retroviruses. Cold Spring Harbor Laboratory Press, New Yor, pp 843.

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

Eickbush TH and Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221-234.

Graur D and Li W (2000) Fundaments of molecular evolution. 2nd edition, Sinauer Associates, Sunderland, Massachusetts, pp 439.

Grumbling G and Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34:D484-8.

Havecker ER, Gao X and Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.

Huszar T and Imler JL (2008) *Drosophila* viruses and the study of antiviral host-defense. Adv Virus Res 72:227-265.

Kijima TE and Innan H (2010) On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27:896-904.

Krysan DJ, Rockwell NC and Fuller RS (1999) Quantitative characterization of furin specificity. Energetics of substrate discrimination using an internally consistent set of hexapeptidyl methylcoumarinamides. J Biol Chem 274:23229-23234.

Kyte J and Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132.

Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B and Vaury C (2000) Life cycle of an endogenous retrovirus, *ZAM*, in *Drosophila melanogaster*. J Virol 74:10658-10669.

Letunic I, Doerks T and Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37:D229-32.

Llorens C, Muñoz-Pomer A, Bernad L, Botella H and Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41.

Llorens JV, Clark JB, Martínez-Garay I, Soriano S, de Frutos R and Martínez-Sebastián MJ (2008) *Gypsy* endogenous retrovirus maintains potential infectivity in several species of Drosophilids. BMC Evol Biol 8:302.

Ludwig A, Valente VL and Loreto EL (2008) Multiple invasions of Errantivirus in the genus *Drosophila*. Insect Mol Biol 17:113-124.

Malik HS and Henikoff S (2005) Positive selection of *Iris*, a retroviral envelope-derived host gene in *Drosophila melanogaster*. PLoS Genet 1:e44.

Malik HS, Henikoff S and Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307-1318.

Matsuda M, Ng CS, Doi M, Kopp A and Tobari YN (2009) Evolution in the *Drosophila ananassae* species subgroup. Fly 3:157-69.

Mejlumian L, Pélisson A, Bucheton A and Terzian C (2002) Comparative and functional studies of *Drosophila* species invasion by the *gypsy* endogenous retrovirus. Genetics 160:201-209.

Minervini CF, Viggiano L, Caizzi R and Marsano RM (2009) Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene 440:42-49.

Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on *tRNA* availability. J Mol Evol 37:273-280.

Mourier T and Willerslev E (2009) Retrotransposons and non-protein coding RNAs. Brief Funct Genomic Proteomic 8:493-501.

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

Nicholas KB and Nicholas HBJ (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author. http://www.psc.edu/biomed/genedoc.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Ronquist F and Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rozas J, Sanchez-DelBarrio JC, Messeguer X and Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Sharp PM and Li WH (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398-402.

Shields DC, Sharp PM, Higgins DG and Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716.

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

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599.

Varela M, Spencer TE, Palmarini M and Arnaud F (2009) Friendly viruses: the special relationship between endogenous retroviruses and their host. Ann N Y Acad Sci 1178:157-172.

Vidal NM, Ludwig A and Loreto ELS (2009) Evolution of *Tom, 297, 17.6* and *rover* retrotransposons in Drosophilidae species. Mol Gen Genomics 282:351-362.

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913-922.

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23-29.

Xu Z and Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-8

TABLES

Table 1: Distribution and abundance of *Kanga-like* endogenous retroviruses and *Iris* gene in 12 *Drosophila* species

Species	**Number of sequences[1] (full length copies)[2]**			
	Kanga1A	*Kanga1B*	*Kanga2*	*Iris*
D. simulans	-	-	-	1
D. sechellia	-	-	-	1
D. melanogaster	-	-	-	1
D. yakuba	-	-	-	1
D. erecta	-	-	-	1
D. ananassae	10 (5)	4 (3)	24 (6)	1
D. pseudoobscura	11	2	-	1
D. persimilis	34	4 (1)	-	1
D. willistoni	-	16 (8)		-
D. mojavensis	-	4 (1)	-	-
D. virilis	-	-	-	-
D. grimshawi	7 (1)	-	-	-
total	62 (6)	30 (13)	24 (6)	

[1] - The number of sequences does not reflect necessarily the number of copies because some scaffolds are very short and those may be duplicated in the dataset. Also, the searches were based on the *env* gene, thus, degenerated copies without *env* gene were not counted;

[2] - Full length copies contain two LTRs and a central region with *gag*, *pol* and *env* genes with reading frames conserved or not.

Table 2: *Kanga-like* endogenous retroviruses characterized in this study

	Species	TSDs (bp)	LTRs (bp)	PBS	Structure	Furin cleavage site in the Env protein	Total length (bp)	LTR divergence range	Age estimate (million years)
Kanga1A									
	D. ananassae	5	477	Trp	*gag_pol_env*	RQRR	8490	0.004 - 0.021	0.125 - 0.65625
	D.pseudoobscura	?	?	?	*gag_pol_env*	RKRR	?	?	?
	D.psersimilis	?	?	?	degenerated ORFs	RKRR	?	?	?
	D. grimshawi	5	520	Trp	degenerated ORFs	RKRR	8180	0.035	1.09375
Kanga1B									
	D. willistoni	5	467	Trp	*gag_pol_env*	RQRR	8961	0.00 - 0.058	0 - 1.8125
	D. pseudoobscura	?	?	?	*gag_pol_env*	RRQR	?	?	?
	D.persimilis	5	658	?	degenerated ORFs	RRQR	8783	0.008	0.25
	D. ananassae	5	414	Arg	*gag_pol_env*	RKVR	8315	0.00 - 0.005	0 - 0.15625
	D. mojavensis	?	490	Trp	degenerated ORFs	RPKR	8876	0.012	0.375
Kanga2									
	D. ananassae	5	1015	Trp	*gag_pol/env*	RQKR	9877	0.001-0.055	0.03125 - 1.71875

Note: General characteristics are based on a consensus sequence or a representative sequence from each species; In structure: (_) genes are in the same reading frame; (/) genes are in different reading frames.

Table 3: dS values for *Kanga1A* are given below and left of the diagonal line, and dS for *amd* gene are given above and right of the diagonal line.

	D.pseudoobscura	*D.persimilis*	*D.ananassae*	*D.grimshawi*
D.pseudoobscura	-	0.06	0.684	0.715
D. persimilis	0.031	-	0.671	0.703
D. ananassae	0.815	0.815	-	0.712
D. grimshawi	0.743	0.725	0.856	-

Table 4: dS values for *Kanga1B* are given below and left of the diagonal line, and dS for *amd* gene are given above and right of the diagonal line.

	D.willistoni	*D.mojavensis*	*D.pseudoobscura*	*D.persimilis*	*D.ananassae*
D.willistoni	-	0.762	0.880	0.854	0.760
D.mojavensis	0.772	-	0.749	0.740	0.710
D.pseudoobscura	0.762	0.743	-	0.060	0.684
D.persimilis	0.746	0.763	0.105	-	0.671
D.ananassae	0.721	0.722	0.762	0.730	-

Table 5: Codon bias index (CBI) and number of effective codons (Nc) for *amd* gene and *kanga 1A* and *Kanga1B*.

Species	*amd*		*Kanga1A*		*Kanga1B*	
	Nc	CBI	Nc	CBI	Nc	CBI
D. ananassae	43.5	0.50	56.0	0.23	48.0	0.37
D. grimshawi	51.0	0.33	51.0	0.38	-	-
D. mojavensis	52.0	0.34	-	-	51.5	0.33
D. willistoni	46.5	0.37	-	-	53.5	0.24
D. pseudoobscura	50.0	0.38	54.0	0.25	52.5	0.30
D. persimilis	51.5	0.33	54.0	0.25	52.5	0.31

Nc varies between 21 (for maximum bias) and 61 (for minimum bias) and CBI varies between 0 (no bias) and 1 (maximum bias)

FIGURES

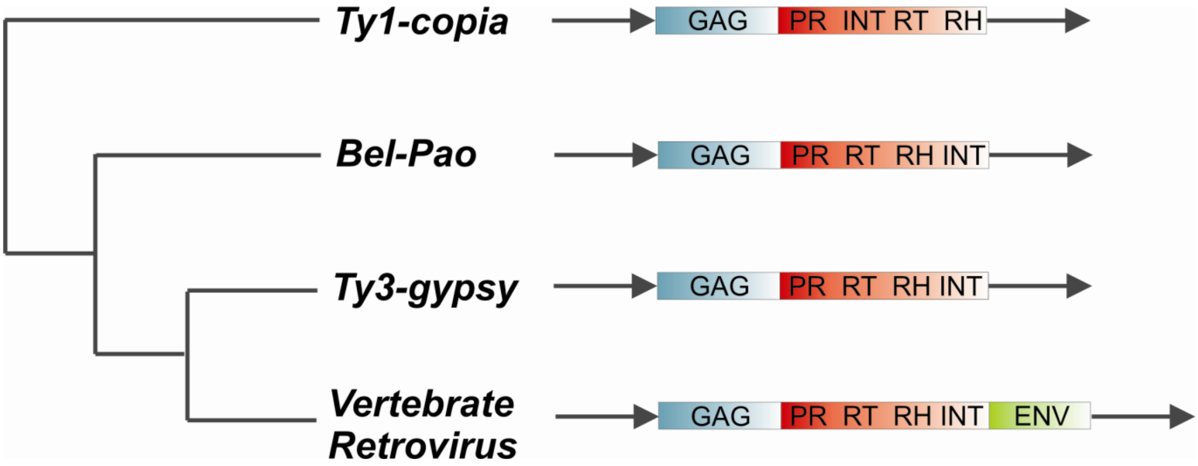


Figure 1: The RT relationships among LTR retroelements and the schematic organization of these sequences. Gray arrows represent the long terminal repeats (LTRs). The *gag* gene is represented by blue box. Red box shows the domains of the gene *pol* that have different order among *Ty1-copia* and the others retroelements. The *env* gene (green box) is present mainly in Retrovirus, but it is also present in some lineages in the other groups (not shown). GAG, Capsid protein; PR, Protease; INT, Integrase; RT, Reverse Transcriptase; RH, RibonucleaseH; ENV, Envelope protein.

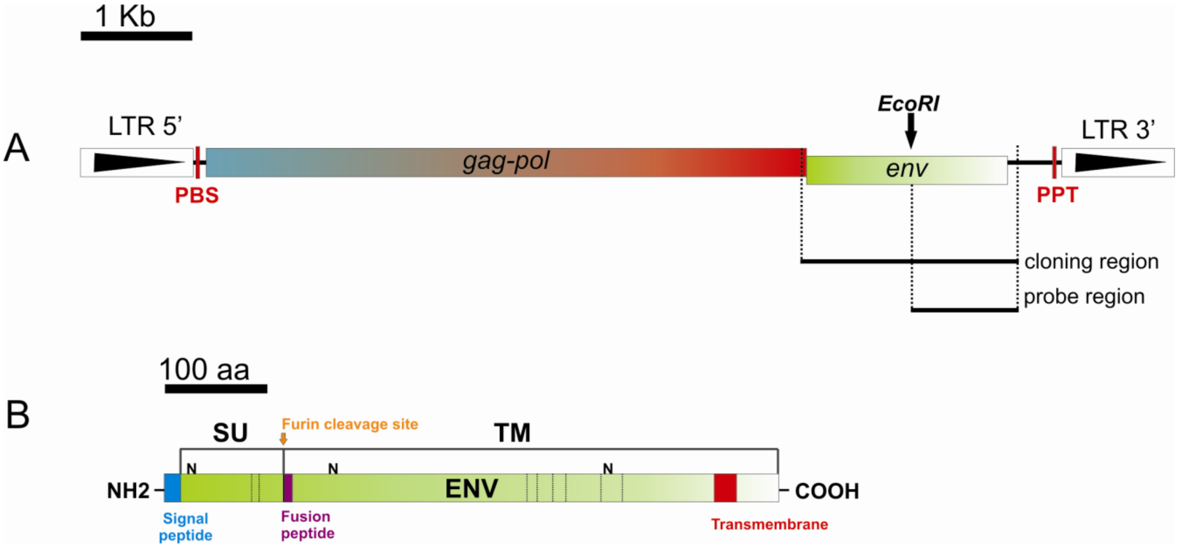


Figure 2: (A) Organization of *Kanga2* retrotransposon composed by two LTRs flanking a central region, which contain *gag*, *pol* and *env* genes. The positions of the primer binding site (PBS), the polypurine tract (PPT), the cloning region, the *EcoRI* site and the probe region are indicated. (B) Schematic *Kanga2* Env protein with the two subdomains, Surface (SU) and Transmembrane (TM). Potential N-glycosylation sites (N), the signal peptide (blue box), the putative furin-like cleavage site (orange arrow), the putative fusion peptide (purple box), the transmembrane peptide (red box), and the cysteine residues (dashed lines) are shown.

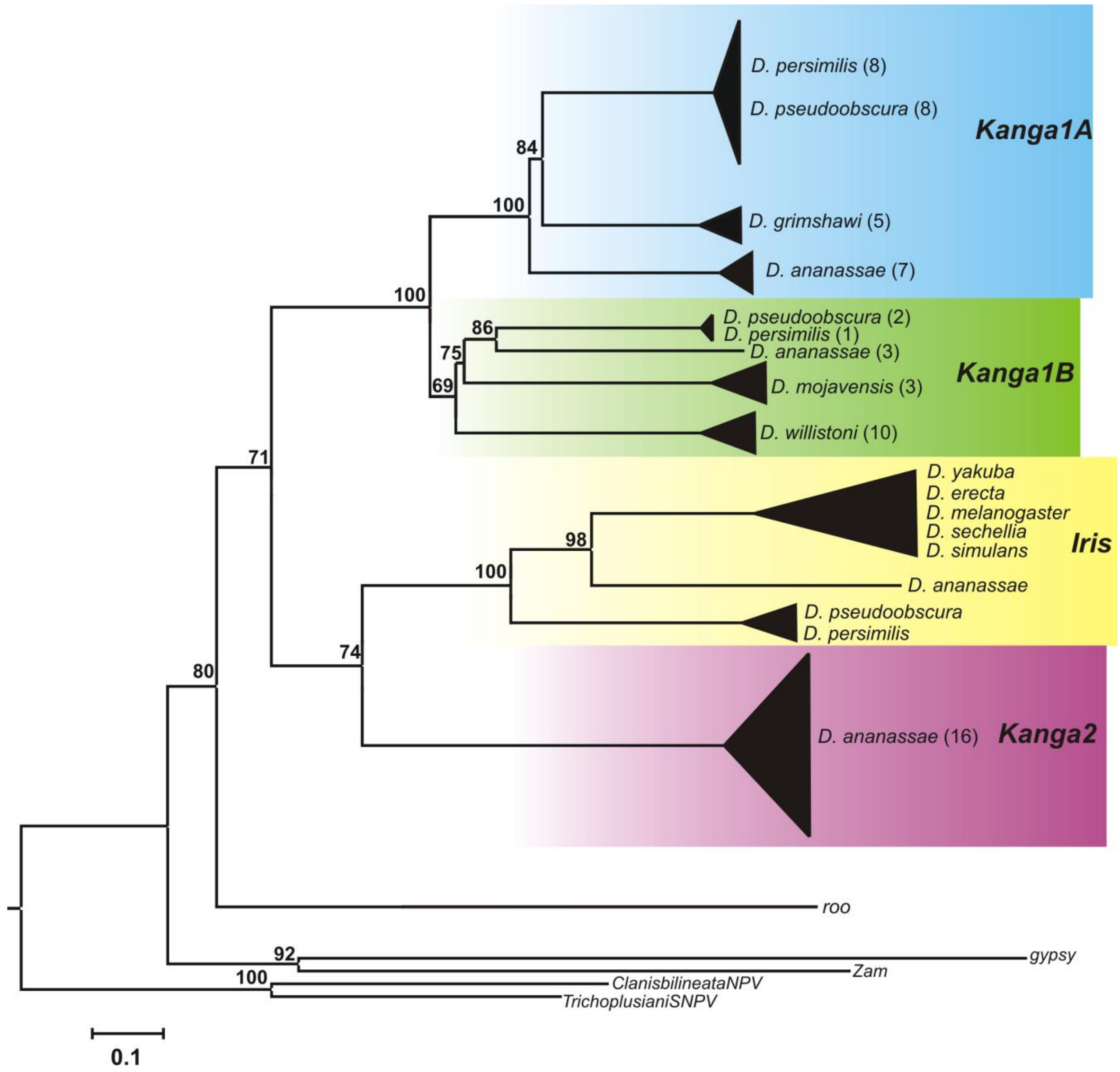


Figure 3: Neighbor-joining tree of *Kanga-like* retrotransposons and *Iris*. The tree is based on the amino acid sequences of the homologous *env* gene. Bootstrap values are shown at the nodes. The number in brackets indicates the number of sequences present in the respective species. Black triangles indicate the divergence and diversity of the sequences within the terminal clades. Envelope sequences of *Trichoplusia ni Single Nuclear Polyhedrosis Virus* (SNPV) and *Clanis bilineata Nucleo Polyhedro Virus* (NPV) were used as outgroup along with sequences of the insect endogenous retroviruses *roo* (from *Bel-Pao* group) and *gypsy* and *Zam* (from *Ty3-gypsy* group).

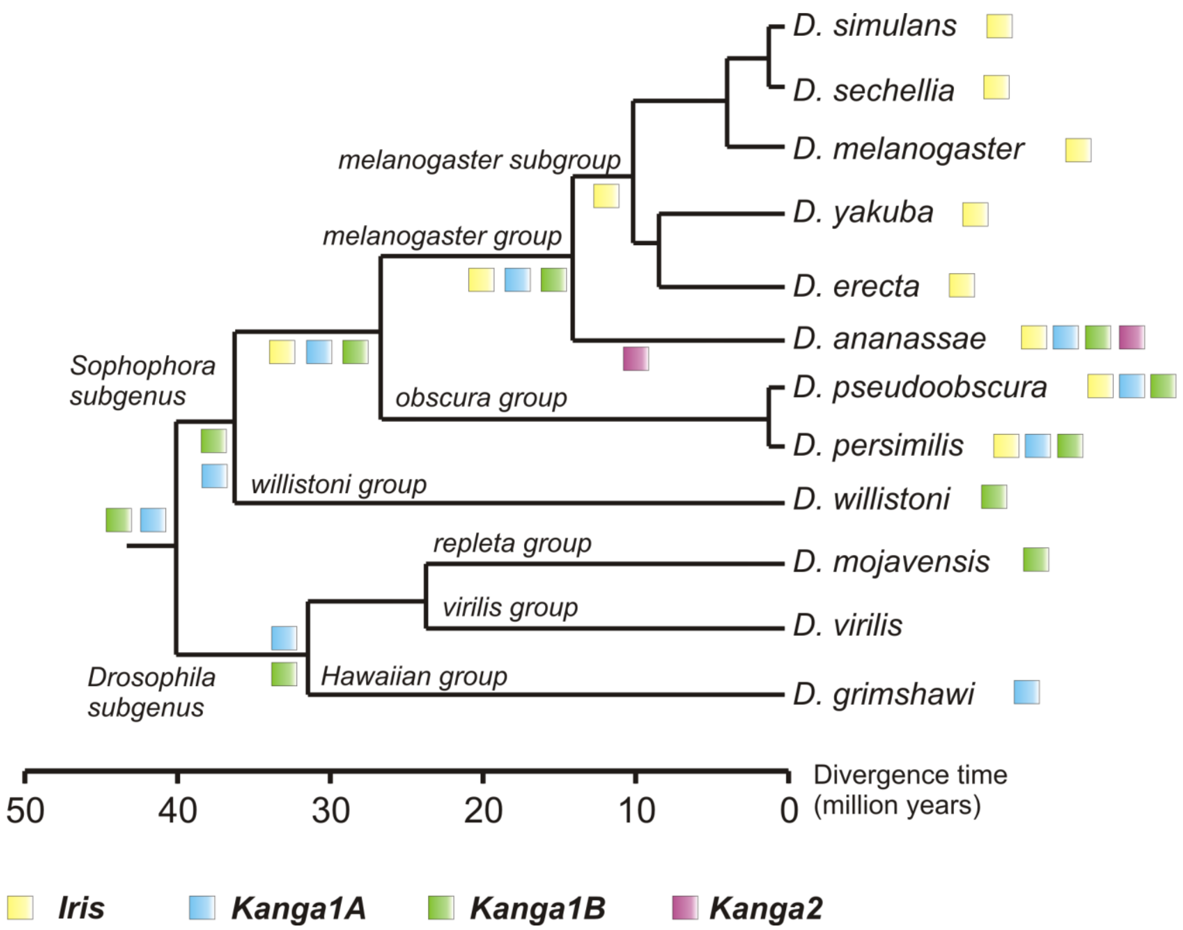


Figure 4: Phylogeny of the 12 sequenced *Drosophila* genomes. Boxes of distinct colors on the right side of species names correspond to the different lineages of *kanga* and *Iris* that are found in each species. The boxes in the internal branches indicated the possible presence of the sequences in the ancestor species.

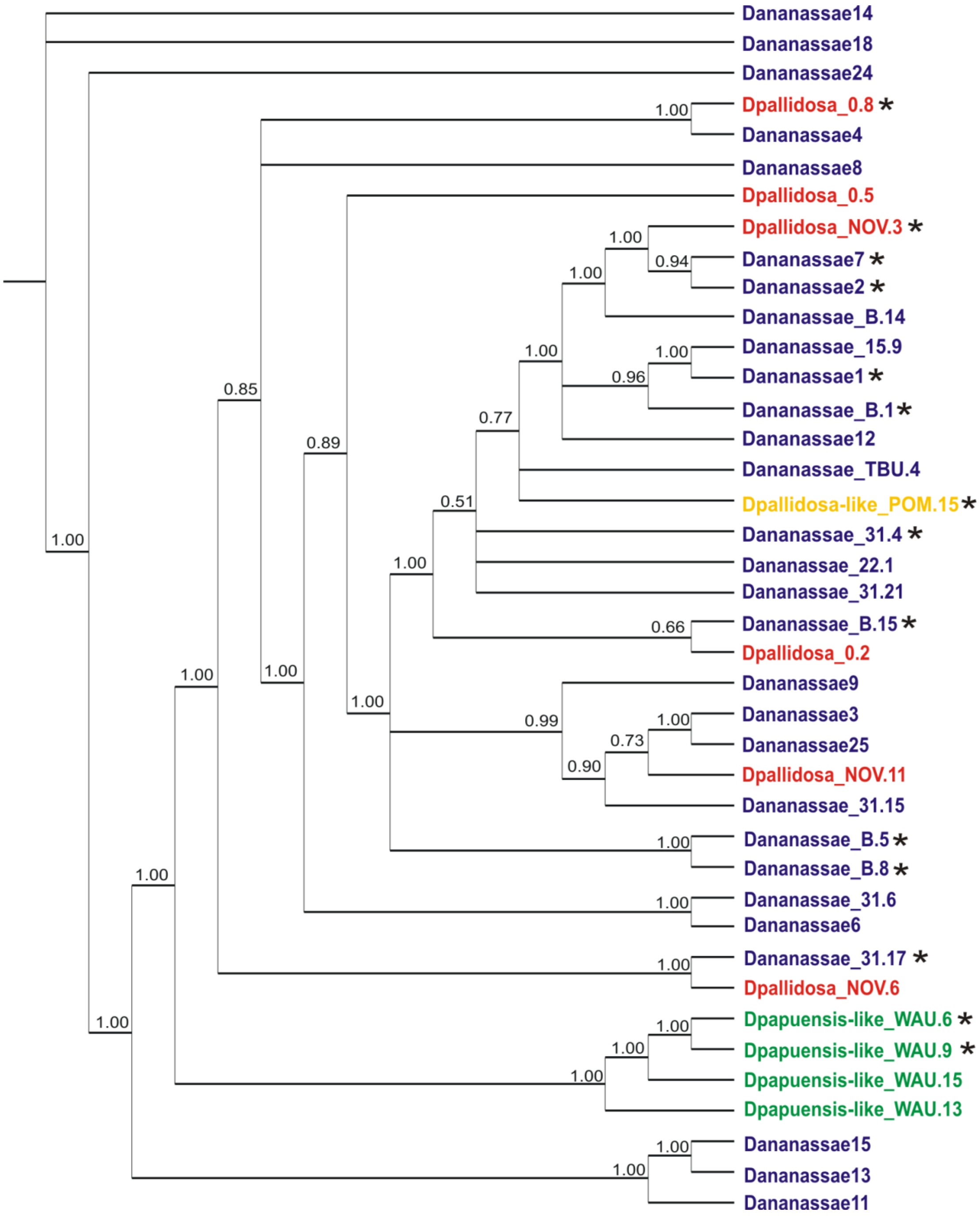


Figure 5: Bayesian phylogenetic tree of *Kanga2* based on nucleotide sequences of *env* gene. Numbers at nodes represent posterior probabilities. Different species are indicated by distinct colors. Sequences putatively encoding Env proteins are marked with asterisk. The remaining sequences present some frameshift or premature stop codon mutation. Sequences from *D. ananassae* genome are followed by a number. Sequences originated by cloning are named as: species name _ strain . number of the clone.

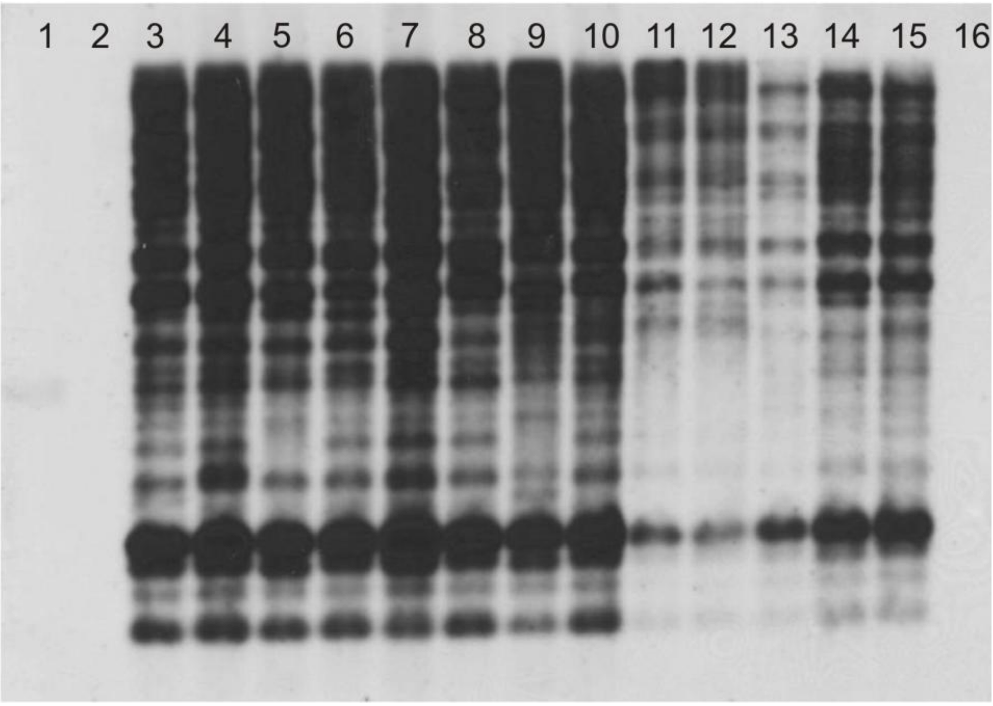


Figure 6: Southern blot of *Kanga2* in different strains of *D. ananassae* and *D. pallidosa.* 1 – 1 Kb ladder; 2 – no DNA; 3 – *D. ananassae18;* 4 – *D. ananassae24;* 5 – *D. ananassae15;* 6 – *D. ananassae23;* 7 – *D. ananassae27;* 8 – *D. ananassae26;* 9 – *D. ananassae35;* 10 – *D. ananassae22;* 11 – *D. ananassae19;* 12 – *D. ananassae23;* 13 – *D. ananassae31;* 14 – *D. pallidosa* (NOV 88); 15 – *D. pallidosa* (VAV 92); 16 – *D. melanogaster* (negative control)

SUPPLEMENTARY MATERIAL

Table S1: Strains used in the PCR screenings.

Genus	Subgenus	Group	Subgroup	Species
Drosophila	*Sophophora*	*melanogaster*	*melanogaster*	*D. melanogaster*
				D. simulans
				D. mauritiana
				D. teissieri
				D. santomea
				D. erecta
				D. yakuba
				D. orena
			montium	*D. lacteicornis*
				D. serrata
				D. auraria
				D. seguyi
				D. biauraria
				D. birchii
			suzukii	*D. suzukii*
				D. lucipennis
			ananassae	*D. ananassae*
				D. pallidosa
				D. pallidosa-like
				D. pallidosa-like Wau
				D. atripex
				D. bipectinata
				D. parabipectinata
				D. ochrogaster
				D. pseudoananassae
				D. malerkotliana
				D. monieri
			takahashii	*D. pseudotakahashii*
				D. takahashii
				D. lutenscens
				D. paralutea
				D. prostipennis
			ficusphila	*D. fiscusphila*
			eugracilis	*D. eugracilis*
			elegans	*D. elegans*
		obscura		*D. affinis*
				D. miranda
				D. subobscura
		saltans		*D. prosaltans*
				D. saltans
				D. sturtevanti
		willistoni	*willistoni*	*D. paulistorum*
				D. willistoni
				D. equinoxialis
				D. tropicalis
				D. capricorni
	Drosophila	*guarani*		*D. ornatifrons*
				D. subbadia
				D. guaru
		guaramunu		*D. griseolineata*
				D. maculifrons
				D. crocina
				D. paramediostriata
		tripunctata		*D. tripunctata*
				D. mediodifusa
				D. mediopictoides
		cardini		*D. cardini*
				D. cardinoides

		D. neocardini
		D. polymorpha
		D. arawacana
	pallidipennis	*D. pallidipennis*
	calloptera	*D. ornatipennis*
	immigrans	*D. immigrans*
		D. albomicans
		D. nasuta
	funebris	*D. funebris*
	mesophragmatica	*D. gasici*
		D. brncici
		D. gaucha
		D. pavani
	repleta	*D. buzzatii*
		D. repleta
	virilis	*D. virilis*
		D. americana
		D. texana
		D. novamexicana
	robusta	*D. robusta*
Zaprionus		*Z. indianus*
		Z. tuberculatus
Scaptodrosophila		*S. lebanonensis*

Capítulo 6

Discussão Geral

Já se passaram muitos anos desde a descoberta dos TEs por Barbara McClintock nos anos 40. Por vários anos, os TEs eram referidos como DNA egoísta e DNA lixo. Apesar da natureza parasítica dos TEs, tentando tirar proveito da maquinaria da célula para se replicar, sua presença nos genomas resulta em uma complexa interação TE-hospedeiro que tem um impacto mensurável sobre a evolução dos mesmos. Não existe mais razão para se referir aos elementos de transposição como DNA lixo, pois um grande número de trabalhos tem mostrado que essas seqüências podem moldar a estrutura, função e regulação dos genomas.

Uma maior compreensão sobre diversidade, evolução, regulação, interação dos elementos móveis está sendo um resultado direto do seqüenciamento de diversos genomas. *Drosophila* tem sido usada, já há muitos anos, como um modelo em várias áreas de estudo e tem proporcionado importantes contribuições para o entendimento dos TEs e seu impacto na evolução genômica.

Essa tese apresenta quatro trabalhos evolutivos de TEs baseados em análises nos 12 genomas de *Drosophila* seqüenciados e em buscas adicionais de seqüências homólogas em um grande número de espécies. Podemos separar os TEs estudados aqui em 3 grupos: elementos *gypsy-like*, também chamados *Errantivirus*, que pertencem ao grupo *Ty3-gypsy* de retrotransposons, apresentados nos Capítulos 2 e 3; o retroelemento *Kanga*, apresentado no Capítulo 5, pertence a um grupo diferente dos anteriores, grupo *Bel-Pao*; e um grupo completamente diferente de transposons de DNA não autônomos, chamados MITEs, abordado no Capítulo 4.

Um dos focos principais dessa tese foi a realização de uma análise evolutiva de retrotransposons no gênero *Drosophila*, com ênfase para a detecção de ocorrência de transferência horizontal e a associação com a capacidade intrínseca desses elementos de serem os próprios vetores dessa transferência.

Através de buscas *in silico* nos 12 genomas de *Drosophila* disponíveis, juntamente com clonagem e seqüenciamentos adicionais, realizamos um amplo estudo evolutivo dos retrotransposons *gypsy, gypsy2, gypsy3, gypsy4* e *gypsy6* no gênero *Drosophila,* apresentadas no Capítulo 1. Evidenciamos um padrão muito complexo de evolução dos elementos estudados, com polimorfismo ancestral, perda estocástica e TH. Nesse trabalho, nós estudamos e desenvolvemos uma metodologia que pode ser empregada para ajudar nas inferências de transferência horizontal (TH). A comparação de valores dS foi escolhida para a inferência de TH porque é uma medida de evolução neutra na ausência de forte uso tendencioso de códons. Assim, seria esperada a mesma proporção de substituições entre TEs e genes nucleares ao comparar duas espécies. No entanto, existe uma diferença na intensidade de seleção no uso de códons sinônimos entre genes, o que influencia na magnitude de dS entre os mesmos por uma correlação negativa (Shields et al. 1988; Sharp e Li 1989). Inicialmente o gene *amd* foi escolhido para as comparações por possuir o maior número de seqüências disponíveis para as espécies, pois estávamos analisando um grande número de seqüências de *gypsy* obtidas no GenBank, de espécies sem genoma sequenciado. Estimamos o nível de uso preferencial de códons para o gene *amd* e para os retroelementos e observamos que este gene é bastante útil para estas comparações, já que não apresenta um grau muito elevado de uso preferencial de códons, o que levaria a valores de dS muito baixos e dificultaria a detecção de casos de TH. Além disso, este gene possui um menor grau de uso preferencial de códons do que os retrotransposons, resultando em valores de dS esperados para o gene menores do que para os TEs. Assim, se um TE foi transmitido de uma espécie para outra por TH, é esperado que o dS do TE entre as espécies seja significativamente menor que os dS encontrado para genes nucleares, refletindo o menor tempo de divergência das seqüências em relação a seqüência ancestral. Dependendo do tamanho da seqüência analisada pode-se utilizar um teste de qui-quadrado ou teste exato de Fisher para acessar a significância das diferenças de dS.

O retrotransposon *nik*, também conhecido como *gypsy5*, foi analisado separadamente no Capítulo 3, pois é um elemento bastante divergente do clado *gypsy*, como pode ser visto na Figura 3 deste capítulo. Nós realizamos buscas *in silico* por elementos completos nos doze genomas de *Drosophila* disponíveis e buscamos por seqüências homologas ao gene *env* por PCR em um grande número de espécies. A presença de cópias distantes e degeneradas sugere que *nik* é antigo nos genomas de *Drosophila.* Os mesmos processos evolutivos identificados no trabalho anterior também devem ter atuado sobre esse elemento e possivelmente são constantes na evolução de TEs, como polimorfismo ancestral, perda estocástica, introgressão e transferência horizontal. Nós também buscamos informações sobre a regulação do elemento *nik*. Existe um mecanismo geral de silenciamento dos retrotransposons de *Drosophila*, que provavelmente controla também o retrotransposon *nik.* Os elementos *gypsy-like* são expressos principalmente nas células somáticas gonadais (células foliculares) e podem infectar as células germinativas por partículas virais produzidas. Um mecanismo distinto de regulação desses retroelementos em relação aos demais transposons foi apontado por Malone et al. (2009). Porém, ambos os mecanismos envolvem a produção de piRNAs que silenciam seqüencias alvo. O lócus produtor de piRNAs *flamenco* foi apontado como a principal fonte de piRNAs atuando nas células somáticas para regulação dos retrotransposons. Nós analisamos então as regiões de *nik* presentes no lócus *flamenco* de três espécies (para as quais os lócus foram identificados), *D. melanogaster, D. yakuba* e *D. erecta.* Ao contrário de *D. yakuba*, que possui várias regiões de *nik*, em *D. melanogaster* e *D. erecta flamenco* pode não ser o principal regulador, pois apenas pequenas regiões de *nik* são encontradas. Nós sugerimos que alguns componentes deste processo de regulação pode ser diferente entre as espécies como um resultado da coevolução de diferentes lócus de piRNAs e dos retrotransposons presentes no genoma.

Nossas análises nos Capítulos 2 e 3 sugerem inúmeros casos de TH de retrotransposons. Grande parte desses casos aconteceu entre as espécies do subgrupo *melanogaster.* Existe um aumento no número de casos descritos de TH entre espécies deste subgrupo (Ludwig e Loreto 2007; Vidal et al. 2009; Sánchez-Gracia et al. 2005; Bartolomé et al. 2009). Esse fato levanta a questão sobre o possível efeito de introgressão na aquisição de novas seqüências nessas espécies. Eventos de introgressão de TEs podem ter acontecido principalmente entre *D*. *melanogaster*, *D*. *simulans* e *D. sechellia*, ou entre *D. santomea, D. yakuba* e *D. teissieri*, ou entre *D. erecta* e *D. orena*. Para as espécies *D*. *melanogaster*, *D*. *simulans* e *D. sechellia* tem sido observada a produção de híbridos férteis dependendo da combinação das linhagens parentais (Coyne 1985; Davis et al. 1996; Sawamura et al. 2000).

Outra importante particularidade dos eventos de TH evidenciados é a possibilidade de os retrotransposons analisados serem agentes infecciosos, o que poderia ajudar a explicar o grande número de THs. De maneira geral, nosso trabalho sugere que várias famílias de retrovírus endógenos têm conservado o gene *env* e poderiam ainda, ou num passado recente, ter propriedades infecciosas.

No Capítulo 5 apresentamos uma análise evolutiva dos retrovírus endógenos *Kanga-like.* Essa linhagem foi identificada no genoma de *D. ananassae* como sendo a provável origem de um gene de envelope viral domesticado em *Drosophila*, *Iris* (Malik e Henikoff 2005). *Kanga* é proximamente relacionado ao elemento *roo*, e interessantemente, apesar de pertencerem a uma linhagem evolutiva independente dos elementos *gypsy-like* (ver Figura 3 do Capítulo 1) esses retroelementos adquiriram independentemente o gene *env* de uma mesma fonte, os baculovírus. Em nossas análises confirmamos a origem de *Iris* no ancestral dos grupos *melanogaster* e *obscura,* cerca de 25 milhões de anos atrás. Encontramos três linhagens antigas de retrovírus *Kanga* que em algumas espécies de *Drosophila*, ainda parecem ter cópias ativas. A linhagem *Kanga2*, que é a mais próxima do gene *Iris*, foi mais intensivamente analisada e mostrou uma distribuição muito restrita. A proteína Env predita de *Kanga2* possui as características que conferem propriedade fusogênica, que é necessária para a infecciosidade. Nós evidenciamos que os retrovírus endógenos *Kanga-like* foram transmitidos verticalmente por um longo período evolutivo nos genomas de *Drosophila*, sendo um modelo importante de co-evolução com uma relação mutualística entre elementos transponíveis e seu hospedeiro.

Apesar da sua capacidade de proliferação, os TEs são freqüentemente perdidos nas espécies hospedeiras, possivelmente passando por um processo de degeneração, como podemos observar nos retrotransposons que estudamos.

Principalmente as análises de seqüências completas de *nik* e *Kanga* revelam diferentes estágios de degeneração e perda completa em algumas espécies. Isso reforça a importância de TH para a sobrevivência dos TEs, um mecanismo que permite um TE invadir um novo genoma que ainda não possui defesa contra ele.

Um tipo interessante de TEs, especialmente estudados e abundantes em plantas, são os MITEs. MITEs são elementos curtos com características peculiares (ver seção 4 do Capítulo1). Relativamente poucos MITEs têm sido descobertos em *Drosophila*. O Capítulo 4 desta tese apresenta um estudo sobre o elemento *Mar*, um MITE. Obtivemos amplificação de fragmentos do elemento somente nas espécies do subgrupo *willistoni* investigadas: *D. willistoni*, *D. paulistorum, D. equinoxialis, D. tropicalis* e *D. insularis.* Os amplicons de cada espécie foram clonados e sequenciados para a realização de análises evolutivas do elemento. No entanto, ainda não foram obtidas as seqüencias dos clones de *D. tropicallis*, para os quais *primers* internos foram desenhados para permitir o seqüenciamento completo dos clones de aproximadamente 3 Kb. Adicionalmente às análises por PCR, realizamos buscas *in silico* nos 12 genomas de *Drosophila*. Como esperado, o elemento *Mar* foi encontrado no genoma de *D. willistoni*, porém, nenhuma cópia foi encontrada nos outros genomas. Encontramos cerca de 80 cópias completas (TIRs intactas) deste elemento no genoma de *D. willistoni,* sendo que a maior parte dessas sequências também apresenta TSDs intactas indicando possível mobilização recente desse elemento. Consistente com prévios trabalhos mostrando associação entre MITEs e genes, nós encontramos 32 cópias do elemento *Mar* localizados perto ou dentro de genes. Em nossas buscas, nenhuma cópia do elemento *Mar* é codificadora de transposase, porém citamos possíveis elementos que poderiam produzir transposase capaz de mobilizar essas seqüências. Ainda, pela comparação da filogenia obtida para o elemento *Mar* com a filogenia das espécies hospedeiras e a observação de alta divergência do elemento, podemos sugerir que a origem deste MITE ocorreu após a separação dos subgrupos *willistoni* e *bocainensis* (desde que espécies deste último subgrupo não possuem o elemento), porém antes do início da diversificação do subgrupo *willistoni* estimado em 5,6 Mya. Uma grande amplificação dessas seqüências ocorreu posteriormente, pelo menos no genoma de *D. willistoni*. Uma análise por southern blot também deverá ser feita para a preparação do manuscrito final. Queremos analisar se as outras espécies do grupo também apresentam um grande número de cópias como *D. willistoni*.

Nosso trabalho visou contribuir para o entendimento da dinâmica evolutiva, diversidade e implicações dos TEs nos genomas de *Drosophila.* A persistência de um TE nos genomas deve ser resultado de influências de seleção positiva, negativa e neutra, que permitem um balanço entre amplificação, excisão, degeneração, e invasão por HT.

Referências Bibliográficas

Arensburger P, Kim YJ, Orsetti J, Aluvihare C, O'Brochta DA e Atkinson PW (2005) An active transposable element, *Herves*, from the African malaria mosquito *Anopheles gambiae*. Genetics 169:697-708.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, Pachter L, Myers E e Langley CH (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310.

Biémont C e Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443:521-524.

Boeke JD e Stoye JP (1997) Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin JM, Hughes SH e Varmus HE (eds) Retroviruses. Cold Spring Harbor Laboratory Press, New York, pp 343-435.

Boussy IA e Daniels SB (1991) *hobo* transposable elements in *Drosophila melanogaster* and *D. simulans*. Genet Res 58:27-34.

Bowen NJ e McDonald JF (2001) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11:1527-1540.

Bowen NJ e Jordan IK (2002) Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol 4:65-76.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R eHannon GJ (2007) Discrete small RNA-generating *loci* as master regulators of transposon activity in *Drosophila*. Cell 128:1089-1103.

Britten RJ (1996). DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A 93:9374-9377.

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

Bureau TE e Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283-94.

Cáceres M, Ranz JM, Barbadilla A, Long M e Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. Science 285:415-418.

Calvi BR, Hong TJ, Findley SD e Gelbart WM (1991) Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: *hobo*, *Activator*, and *Tam3*. Cell 66:465-471.

Capy P, Bazin C, Higuet D and Langin T (1998) Dynamics and evolution of transposable elements. Landes Bioscience, Austin, Texas, 197 pp.

Carmell MA, Xuan Z, Zhang MQ e Hannon GJ (2002) The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev 16:2733-2742.

Casals F, Cáceres M e Ruiz A (2003) The *foldback*-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. Mol Biol Evol 20:674-685.

Chalvet F, Teysset L, Terzian C, Prud'homme N, Santamaria P, Bucheton A e Pélisson A (1999) Proviral amplification of the *Gypsy* endogenous retrovirus of *Drosophila melanogaster* involves *env*-independent invasion of the female germline. EMBO J 18:2659-2669.

Chung WJ, Okamura K, Martin R e Lai EC (2008) Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. Curr Biol 18:795-802.

Coates CJ, Johnson KN, Perkins HD, Howells AJ, O'Brochta DA e Atkinson PW (1996) The *hermit* transposable element of the Australian sheep blowfly, *Lucilia cuprina*, belongs to the *hAT* family of transposable elements. Genetica 97:23-31.

Coffin JM, Hughes SH and Varmus HE (1997) Retroviruses. Cold Spring Harbor Laboratory Press, New Yor, pp 843.

Cordeiro AR and Winge H (1995) Levels of evolutionary divergence of *Drosophila willistoni* sibling species. In: Levine L (ed) Genetics of natural populations: the continuing importance of Theodosius Dobzhansky. Columbia University Press, New York, pp 262-280.

Coyne JA (1985) Genetic studies of three sibling species of *Drosophila* with relationship to theories of speciation. Genet Res 46:169-192.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004) WebLogo: A sequence logo generator. Genome Research 14:1188-1190.

Daniels SB, Chovnick A e Boussy IA (1990a) Distribution of *hobo* transposable elements in the genus *Drosophila*. Mol Biol Evol 7:589-606.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG e Chovnick A (1990b) Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics 124:339-355.

Davis AW e Wu CI (1996) The broom of the sorcerer's apprentice: the fine structure of a chromosomal region causing reproductive isolation between two sibling species of *Drosophila*. Genetics. 143:1287-1298.

de la Chaux N e Wagner A (2009) Evolutionary dynamics of the LTR retrotransposons *roo* and *rooA* inferred from twelve complete *Drosophila* genomes. BMC Evol Biol 9:205.

Delprat A, Negre B, Puig M e Ruiz A (2009) The transposon *Galileo* generates natural chromosomal inversions in *Drosophila* by ectopic recombination. PLoS One 4:e7883.

Deprá M, Panzera Y, Ludwig A, Valente VL e Loreto EL (2010) *hosimary*: a new *hAT* transposon group involved in horizontal transfer. Mol Genet Genomics (in press).

Deprá M, Valente VL, Margis R e Loreto EL (2009) The *hobo* transposon and *hobo*-related elements are expressed as developmental genes in *Drosophila*. Gene 448:57-63.

Desset S, Meignin C, Dastugue B e Vaury C (2003) COM, a heterochromatic locus governing the control of independent endogenous retroviruses from *Drosophila melanogaster*. Genetics 164:501-509.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203-218.

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

Eickbush TH e Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221-234.

Farabaugh PJ (1996) Programmed translational frameshifting. Microbiol Rev 60:103-134.

Feschotte C e Pritham EJ (2007) DNA Transposons and the Evolution of Eukaryotic Genomes. Annu Rev Genet 41:331-368.

Feschotte C, Zhang X e Wessler SR (2002). Miniature inverted repeat transposable elements and their relationship to established DNA transposons. In: Craig NL, Craigie R, Gellert M and Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, DC, pp 1147-1158.

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

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

Gonzáles J e Petrov D (2009) Genetics. MITEs-the ultimate parasites. Science 325:1352-1353.

Graur D e Li W (2000) Fundaments of molecular evolution. 2nd edition, Sinauer Associates, Sunderland, Massachusetts, pp 439.

Grumbling G e Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34:D484-8.

Grzebelus D, Gładysz M, Macko-Podgórni A, Gambin T, Golis B, Rakoczy R e Gambin A (2009) Population dynamics of miniature inverted-repeat transposable elements (MITEs) in *Medicago truncatula*. Gene 448:214-220.

Hammer SE, Strehl S e Hagemann S (2005) Homologs of *Drosophila P* transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human. Mol Biol Evol 22:833-844.

Handler AM e Gomez SP (1997) A new *hobo*, *Ac*, *Tam3* transposable element, *hopper*, from *Bactrocera dorsalis* is distantly related to *hobo* and *Ac*. Gene 185:133-135.

Havecker ER, Gao X e Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225.

Hehl R, Nacken WK, Krause A, Saedler H e Sommer H (1991) Structural analysis of *Tam3*, a transposable element from *Antirrhinum majus*, reveals homologies to the *Ac* element from maize. Plant Mol Biol 16:369-371.

Herédia F, Loreto EL e Valente VL (2004) Complex Evolution of *gypsy* in Drosophilid Species. Mol Biol Evol 21:1831-1842.

Holyoake AJ e Kidwell MG (2003) *Vege* and *Mar*: two novel *hAT* MITE families from *Drosophila willistoni*. Mol Biol Evol 20:163-167.

Huszar T and Imler JL (2008) *Drosophila* viruses and the study of antiviral host-defense. Adv Virus Res 72:227-265.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR e Wessler SR (2003) An active DNA transposon family in rice. Nature 421:163-167.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3:RESEARCH0084.

Kapitonov VV e Jurka J (2000) GYPSY5_I. Direct Submission to Repbase Update (SEP-2000).

Kapitonov VV e Jurka J (2002) POGON1, a *bona fide* family of nonautonomous DNA. Repbase Reports 2:7.

Kapitonov VV e Jurka J (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A 100:6569-6574.

Kapitonov VV e Jurka J (2007) *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521-529.

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626-1632.

Kempken F e Windhofer F (2001) The *hAT* family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110:1-9.

Kidwell MG e Lisch DR (2001) Transposable elements, parasitic DNA, and genome evolution. Evolution 55:1-24.

Kijima TE e Innan H (2010) On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27:896-904.

Kim A, Terzian C, Santamaria P, Pélisson A, Purd'homme N e Bucheton A (1994) Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc Natl Acad Sci USA 91:1285-1289.

Koga A, Shimada A, Shima A, Sakaizumi M, Tachida H e Hori H (2000) Evidence for recent invasion of the medaka fish genome by the *Tol2* transposable element. Genetics 155:273-281.

Kohany O, Gentles AJ, Hankus L e Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

Krysan DJ, Rockwell NC e Fuller RS (1999) Quantitative characterization of furin specificity. Energetics of substrate discrimination using an internally consistent set of hexapeptidyl methylcoumarinamides. J Biol Chem 274:23229-23234.

Kyte J e Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B e Vaury C (2000) Life cycle of an endogenous retrovirus, *ZAM*, in *Drosophila melanogaster*. J Virol 74:10658-10669.

Letunic I, Doerks T e Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37:D229-32.

Lewis RL, Beckenbach AT e Mooers AØ (2005) The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on *COI* and *COII* sequences and a Bayesian estimate of phylogeny. Mol Phylogenet Evol 37:15-24.

Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L e Hannon GJ (2004) Argonaute2 is the catalytic engine of mammalian RNAi. Science 305:1437-1441.

Llorens C, Muñoz-Pomer A, Bernad L, Botella H e Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41.

Llorens JV, Clark JB, Martínez-Garay I, Soriano S, de Frutos R e Martínez-Sebastián MJ (2008) *Gypsy* endogenous retrovirus maintains potential infectivity in several species of Drosophilids. BMC Evol Biol 8:302.

Locke J, Howard LT, Aippersbach N, Podemski L e Hodgetts RB (1999) The characterization of *DINE-1*, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. Chromosoma 108:356-366.

Loreto EL, Valente VL, Zaha A, Silva JC e Kidwell MG (2001) *Drosophila mediopunctata P* elements: a new example of horizontal transfer. J Hered 92:375-381.

Loreto EL, Zaha A, Nichols C, Pollock JA e Valente VLS (1998) Characterization of a hypermutable strain of *Drosophila simulans*. Cell Mol Life Sci 54:1283-1290.

Loreto ELS, Carareto CMA e Capy P (2008) Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100:545-554.

Ludwig A e Loreto EL (2007) Evolutionary pattern of the gtwin retrotransposon in the *Drosophila* melanogaster subgroup. Genetica 130:161-8.

Ludwig A e Loreto EL (2007) Evolutionary pattern of the gtwin retrotransposon in the *Drosophila* melanogaster subgroup. Genetica 130(2):161-8.

Malik HS e Henikoff S (2005) Positive selection of *Iris*, a retroviral envelope-derived host gene in *Drosophila melanogaster*. PLoS Genet 1:e44.

Malik HS, Henikoff S e Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307-1318.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R e Hannon GJ (2009) Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. Cell 137:522-535.

Maruyama K e Hartl DL (1991) Evidence for interspecific transfer of the transposable element *mariner* between *Drosophila* and *Zaprionus*. J Mol Evol 33:514-524.

Matsuda M, Ng CS, Doi M, Kopp A e Tobari YN (2009) Evolution in the *Drosophila ananassae* species subgroup. Fly 3:157-69.

McClintock B (1947) Cytogenetic studies of maize and *Neurospora*. Carnegie Institution of Washington. Year Book #46:146-152.

McClintock B (1984) The significance of responses of the genome to challenge. Science. 1984 Nov 16;226(4676):792-801.

McDonald JF (1993) Evolution and consequences of transposable elements. Curr Opin Genet Dev 3:855-864.

Mejlumian L, Pélisson A, Bucheton A e Terzian C (2002) Comparative and functional studies of *Drosophila* species invasion by the *gypsy* endogenous retrovirus. Genetics 160:201-209.

Minervini CF, Viggiano L, Caizzi R e Marsano RM (2009) Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene 440:42-49.

Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on *tRNA* availability. J Mol Evol 37:273-280.

Mota NR, Ludwig A, da Silva Valente VL e Loreto EL (2010) *harrow*: new *Drosophila hAT* transposons involved in horizontal transfer. Insect Mol Biol (in press).

Mourier T e Willerslev E (2009) Retrotransposons and non-protein coding RNAs. Brief Funct Genomic Proteomic 8:493-501.

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

Nekrutenko A e Li WH (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res 10:1986-1995.

Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D e Seoighe C (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J 5:160.

Nicholas KB e Nicholas HBJ (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author. http://www.psc.edu/biomed/genedoc.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Obbard DJ, Gordon KH, Buck AH e Jiggins FM (2009) The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc Lond B Biol Sci 364:99-115.

Oliver KR e Greene WK (2009) Transposable elements: powerful facilitators of evolution. Bioessays 31:703-714.

Ortiz MF and Loreto ELS (2008) The *hobo*-related elements in the *melanogaster* species group. Genet Res 90:243-252.

Ortiz MF and Loreto ELS (2009) Characterization of new *hAT* transposable elements in 12 *Drosophila* genomes. Genetica 135:67-75.

Ortiz MF, Lorenzatto KR, Corrêa BR e Loreto ELS (2010) *hAT* transposable elements and their derivatives: an analysis in the 12 *Drosophila* genomes. Genetica (in press).

Pascual L e Periquet G (1991) Distribution of *hobo* transposable elements in natural populations of *Drosophila melanogaster*. Mol Biol Evol 8:282-296.

Pélisson A, Song SU, Prud'homme N, Smith PA, Bucheton A e Corces VG (1994) *Gypsy* transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the *Drosophila flamenco* gene. EMBO J 13:4401-4411.

Pinkerton AC, Whyard S, Mende HA, Coates CJ, O'Brochta DA e Atkinson PW (1999) The Queensland fruit fly, *Bactrocera tryoni*, contains multiple members of the *hAT* family of transposable elements. Insect Mol Biol 8:423-434.

Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. J Hered 100:648-655.

Prud'homme N, Gans M, Masson M, Terzian C e Bucheton A (1995) *Flamenco*, a gene controlling the *gypsy* retrovirus of *Drosophila melanogaster*. Genetics 139:697-711.

Quesneville H, Nouaud D e Anxolabéhère D (2006) *P* elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. BMC Genomics 7:214.

Ray DA, Pagan HJ, Thompson ML e Stevens RD (2007) Bats with *hATs*: evidence for recent DNA transposon activity in genus *Myotis*. Mol Biol Evol 24:632-639.

Robe LJ, Cordeiro J, Loreto EL e Valente VL (2010) Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila williston*i subgroup (Diptera: Drosophilidae). Genetica (in press).

Robert CE (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Ronquist F e Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rozas J, Sanchez-DelBarrio JC, Messeguer X e Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Rubin E, Lithwick G e Levy AA (2001) Structure and evolution of the *hAT* transposon superfamily. Genetics 158:949-957.

Rubin E, Lithwick G e Levy AA (2001) Structure and evolution of the *hAT* transposon superfamily. Genetics 158:949-957.

Saitou N e Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Sánchez-Gracia A, Maside X e Charlesworth B (2005) High rate of horizontal transfer of transposable elements in *Drosophila*. Trends Genet 21:200-203.

Sarot E, Payen-Groschêne G, Bucheton A e Pélisson A (2004) Evidence for a piwi-dependent RNA silencing of the *gypsy* endogenous retrovirus by the *Drosophila melanogaster flamenco* gene. Genetics 166:1313-1321.

Sassi AK, Herédia FO, Loreto ELS, Valente VLS e Rohde C (2005) Transposable elements *P* and *gypsy* in natural populations of *Drosophila willistoni*. Genet Mol Biol 28:734-739.

Sawamura K, Davis AW e Wu CI (2000) Genetic analysis of speciation by means of introgression into *Drosophila melanogaster*. Proc Natl Acad Sci USA 97(6):2652-2655.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115.

Sharp PM e Li WH (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398-402.

Shields DC, Sharp PM, Higgins DG e Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704-716.

Silva JC e Kidwell MG (2000) Horizontal transfer and selection in the evolution of *P* elements. Mol Biol Evol 17:1542-1557.

Silva JC, Loreto EL e Clark JB (2004) Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol 6:57-71.

Simmons GM (1992) Horizontal transfer of *hobo* transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. Mol Biol Evol 9:1050-1060.

Song SU, Gerasimova T, Kurkulos M, Boeke JD e Corces VG (1994) An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev 8:2046-2057.

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

Subramanian RA, Arensburger P, Atkinson PW e O'Brochta DA (207) Transposable element dynamics of the *hAT* element *Herves* in the human malaria vector *Anopheles gambiae s.s.* Genetics 176:2477-2487.

Sundararajan P, Atkinson PW and O'Brochta DA (1999) Transposable element interactions in insects: crossmobilization of *hobo* and *Hermes*. Insect Mol Biol 8:359-368.

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599.

Tamura K, Subramanian S e Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

Terzian C, Ferraz C, Demaille J e Bucheton A (2000) Evolution of the *Gypsy* endogenous retrovirus in the *Drosophila melanogaster* subgroup. Mol Biol Evol 17:908-914.

Terzian C, Pélisson A e Bucheton A (2001) Evolution and phylogeny of insect endogenous retroviruses. BMC Evol Biol 1:3.

Thornburg BG, Gotea V e Makayowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104-110.

Torres FP, Fonte LFM, Valente VLS e Loreto ELS (2006) Mobilization of a *hobo*-related sequence in the genome of *Drosophila simulans*. Genetica 123:101-110.

Varela M, Spencer TE, Palmarini M e Arnaud F (2009) Friendly viruses: the special relationship between endogenous retroviruses and their host. Ann N Y Acad Sci 1178:157-172.

Vidal NM, Ludwig A e Loreto ELS (2009) Evolution of *Tom, 297, 17.6* and *rover* retrotransposons in Drosophilidae species. Mol Gen Genomics 282:351-362.

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913-922.

Warren WD, Atkinson PW e O'Brochta DA (1994) The *Hermes* transposable element from the house fly, *Musca domestica*, is a short inverted repeat-type element of the *hobo*, *Ac*, and *Tam3* (*hAT*) element family. Genet Res 64:87-97.

Weiss RB, Dunn DM, Shuh M, Atkins JF e Gesteland RF (1989) *E. coli* ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. New Biol 1:159-169.

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A 103:17600-17601.

Wessler SR, Bureau TE e White SE (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814-821.

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

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23-29.

Xing Y e Lee C (2006) Can RNA selection pressure distort the measurement of Ka/Ks? Gene 29:1-5.

Xiong Y e Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362. Bartolomé C, Bello X and Maside X (2009) Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. Genome Biol 10:R22.

Xu J, Wang M, Zhang X, Tang F, Pan G e Zhou Z (2010) Identification of NbME MITE families: potential molecular markers in the microsporidia *Nosema bombycis*. J Invertebr Pathol 103:48-52.

Xu Z e Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-8

Yang G, Nagel DH, Feschotte C, Hancock CN e Wessler SR (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. Science 325:1391-1394.

Yang HP e Barbash DA (2008) Abundant and species-specific *DINE-1* transposable elements in 12 *Drosophila* genomes. Genome Biol 9:R39.

Yang HP, Hung TL, You TL, Yang TH (2006) Genome-wide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. Genetics 173:189-196.

Anexos

IMB_787_sm_Table 1S: List of retroelements sequences obtained from GenBank and Repbase, with the respective accession number and retroelements sequences obtained by searches in the genomes, with the queries sequences and chromosomal localization.

Species	*Nomenclature*	*GenBank*	*Genomes Search*	
		Accession number	Query	Chromosome/ Scaffold/ Trace
D. annulimana	annulimana39	AF548168		
D. bandeirantorum	bandeirantorum3	AF548150		
D. busckii	busckii4	AF548172		
D. erecta	erecta1	AJ308090		
	erecta2		melanogaster1	4929
	erecta3		mediopicta35	4709
	erecta4		melanogaster1	4784
	erecta5		melanogaster1	4929
	erecta6		melanogaser1	3160
	erecta7		erecta1	4929
	erecta8		erecta1	4845
	erecta9		gypsy2	4929
	erecta10		gypsy3	4845
	erecta11		mediopicta35	4929
	erecta12		simulans11	4929
	erecta13		gypsy2	4929
	erecta14		mediopicta35	4724
	erecta15		gypsy2	4929
	erecta16		mediopicta35	4929
	erecta17		melanogaster7	4929
	erecta18		gypsy6	4784
	erecta19		gypsy6	4845
D. griseolineata	griseolineata10	AF548182		
D. hydei	hydei1	AF548148		
D. maculifrons	maculifrons1	AF548177		
D. mediopicta	mediopicta34	AF548195		
	mediopicta35	AF548196		
D. mediopunctata	mediopunctata45	AF548166		
	mediopunctata49	AF548197		
D. mediosignata	mediosignataA7	AF548193		
D. melanogaster	melanogaster1	M12927		
	melanogaster2		simulans3	X
	melanogaster3		melanogaster1	U
	melanogaster4		melanogaster1	U
	melanogaster5		melanogaster1	3h
	melanogaster6		gypsy2	2h
	melanogaster7		gypsy2	U
	melanogaster8		gypsy2	3L
	melanogaster9		gypsy6	X
	gypsy2	FBti0020264		
	gypsy3	FBti0019930		
	gypsy4	FBti0019254		
	gypsy6	FBgn0063431		
D. mojavensis	mojavensis1		mediopunctata45	6680
D. nebulosa	nebulosa24	AF548187		
D. neocardini	neocardini88	AF548169		
D. ornatifrons	ornatifrons26	AF548185		
D. pallidipennis	pallidipennis1	AF548157		
D. paulistorum	paulistorum3	AF548175		
D. persimilis	persimilis1		subobscura1	super 81
D. polymorpha	polymorpha81	AF548161		
D. pseudoobscura	pseudoobscura1		subobscura1	U
D. sechellia	sechellia1		melanogaster1	super 217
	sechellia2		simulans5	super 51
	sechellia3		simulans10	super 389
	sechellia4		gypsy2	super 83
	sechellia5		gypsy3	super 119
	sechellia6		gypsy4	super 154
D. simulans	simulans1	AF548145		
	simulans2		cordeiroiD	U
	simulans3		melanogaster1	X
	simulans4		melanogaster1	2h
	simulans5		mediopicta35	3h
	simulans6		melanogaster1	2h
	simulans7		gypsy3	2h
	simulans8		gypsy4	X
	simulans9		sechellia4	U
	simulans10		melanogaster1	3h
D. subobscura	subobscura1	X72390		
D. teissieri	teissieri1	AJ308092		
D. virilis	virilis1	AJ308094		
	virilis2	M38438		
	virilis3		hydei1	7678
	virilis4		virilis1	9604
D. willistoni	willistoni1		paulistorum3	gi|111135442|gb|CH963921|
D. yakuba	yakuba1		melanogaster1	U
	yakuba2		melanogaster1	U
	yakuba3		melanogaster1	U
	yakuba4		erecta13	2L
	yakuba5		gypsy3	U
	yakuba6		gypsy2	U
	yakuba7		gypsy6	X
D. zottii	zotti65	AF548163		
S. latifasciaeformis	S.latifasciaeformis1	AF548144		
Z. indianus	Z.indianus1	AF548156		

IMB_787_sm_Table 2S: Accession numbers for the *amd* sequences

Species	*Accession number*
D. ornatifrons	AY699250
D. griseolineata	AY699257
D. maculifrons	[a]
D. mediopuntata	AY699255
D. bandeirantorum	AY699256
D. polymorpha	AY699259
D. neocardini	AY699260
D. virilis	AF293729
D. hydei	AF293712
D. mojavensis	scaffold_6500:6,803,591-6,804,021 [b]
D. incompta	AY699247
D. cordeiroi	EU068743[c]
D. flavopilosa	EU068742[c]
D. willistoni	AF293730
D. paulistorum	AF293719
D. nebulosa	AF293717
D. paulistorum	AF293719
D. simulans	AF293726
D. sechellia	super_7: 2742420-2742850 [b]
D. melanogaster	X04695
D. erecta	AF293708
D. yakuba	chr2R: 5628964-5629394 [b]
D. teissieri	AF293727
D. pseudoobscura	AF293706
D. persimilis	AF293720
D. busckii	AF293707
S. latifasciaeformis	AY699264
Z. indianus	AY699263

[a] – species whose sequence was not yet published and was obtained directly with the author (Lizandra Jaqueline Robe, personal comunication).

[b] – the sequences were obtained by *in silico* searches using the BLAT tool (http://genome.ucsc.edu/cgi-bin/hgBlat).

[c] – The *Amd-un2* and *Amd-bw* primers (Tatarenkov *et al.*, 2001) were used to amplify a fragment of *amd* gene of these two species. DNA sequencing was performed directly using purified amplicons in a MegaBACE 500 automatic sequencer.

Reference:

Tatarenkov, A., Zurovcova, M. and, Ayala, F.J. (2001) *Ddc* and *amd* sequences resolve phylogenetic relationships of *Drosophila*. *Mol Phylogenet Evo* **20**: 321-325.