UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Elementos transponíveis na evolução genômica de espécies incipientes: um estudo no grupo *willistoni* de *Drosophila* (Diptera: Drosophilidae)

HENRIQUE DA ROCHA MOREIRA ANTONIOLLI

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Doutor em Genética e Biologia Molecular.

Orientadora: Profa. Dra. Vera Lúcia da Silva Valente Gaiesky

Coorientadora: Profa. Dra. Maríndia Deprá

Porto Alegre, março de 2024

CIP - Catalogação na Publicação

Antoniolli, Henrique da Rocha Moreira
Elementos transponíveis na evolução genômica de espécies incipientes: um estudo no grupo willistoni de Drosophila (Diptera: Drosophilidae) / Henrique da Rocha Moreira Antoniolli. -- 2024.
127 f.
Orientadora: Vera Lúcia da Silva Valente Gaiesky.

Coorientadora: Maríndia Deprá.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Instituto de Biociências, Programa de Pós-Graduação em Genética e Biologia Molecular, Porto Alegre, BR-RS, 2024.

1. Drosophilidae. 2. Genômica. 3. Mobiloma. I. Gaiesky, Vera Lúcia da Silva Valente, orient. II. Deprá, Maríndia, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Este trabalho foi realizado no Laboratório de *Drosophila*, do Departamento de Genética da Universidade Federal do Rio Grande do Sul, e na Sección Genética Evolutiva, da Facultad de Ciencias da Universidad de la República (Uruguai), com bolsa do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (processo n.º 141319/2020-8) e recursos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

Dedico esta tese à minha mãe, Rosângela, por fazer dos meus sonhos, os seus.

AGRADECIMENTOS

À minha mãe, pelo exemplo, incentivo, suporte incondicional e por ter tornado essa jornada até aqui possível.

Ao meu irmão pelo suporte e pela ajuda em todos os momentos.

À minha querida orientadora, Verinha, eterna inspiração tanto profissional quanto pessoal. Muito obrigado pelos ensinamentos, pela disponibilidade, por acreditar no meu trabalho (embarcando nas minhas ideias) e, principalmente, pelo incentivo. Trabalhar com a senhora é uma honra!

À minha coorientadora, Maríndia, também uma inspiração. Obrigado por abrir as portas do laboratório ainda na época do meu mestrado, também acreditando no meu trabalho e estando pronta para ajudar a qualquer momento desde então.

Para ambas Verinha e Maríndia, obrigado por tornar o ambiente de trabalho e todo o processo de construção da tese uma experiência leve e gratificante! Vocês são um exemplo de como uma orientação deve ser conduzida.

Um agradecimento especial para a Liz que, apesar de não ter me orientado oficialmente nesta tese, continuará para sempre como uma mentora. Muito obrigado por me apresentar ao mundo da Genética e, principalmente, das drosófilas!

Aos meus e minhas colegas da UFRGS – Ane, Léo, Mayara, Natasha, Pedro, Thays, Thamis e Vítor – que, mesmo com a convivência afetada pela pandemia, se fizeram presentes. Um agradecimento especial ao Pedro, meu primeiro coorientado, por ter aceitado os desafios (e também embarcado nas minhas ideias); e à Ane e ao Vítor, que foram um suporte fundamental em momentos difíceis.

Aos meus amigos e amigas de infância – Dani, Gilmar, Lisi, Lilo, Luísa e Paula – que apesar da distância dos últimos 10 anos, se fizeram presentes, dando suporte e comemorando comigo as minhas conquistas. Em especial, à Luísa e à Dani, também pelo suporte em diversos momentos difíceis.

Às minhas amigas que a FURG me presenteou – Ana, Fabi, Ju, Larissa e Manu – pelo apoio fundamental de sempre, e sempre que necessário. Acrescento aqui a Tuca que, de certo modo, também me foi apresentada pela FURG e cuja amizade tem sido fundamental desde então.

Aos professores do PPGBM, pelo conhecimento compartilhado.

Ao Elmo e ao Gabriel, pela prontidão em ajudar e resolver qualquer situação.

À Berê e Helena, pelo apoio na manutenção dos estoques de drosófila e do cafezinho e, principalmente, pelos momentos de amizade e descontração entre um experimento e outro.

Y, por último, pero no menos importante, agradezco a Beatriz y Seba por recibirme como miembro de sus familias en Montevideo mientras realizaba la pasantía y por compartir sus conocimientos. El aprendizaje que tuve con cada uno de ustedes fue más allá de lo profesional. También agradezco a todos en la Udelar, quienes también me recibieron muy bien. A Seba, en particular, gracias por su disponibilidad para resolver dudas en cualquier momento.

"(...) Agradecer os amigos que fiz
E que mantém a coragem de gostar de mim, apesar de mim"

(Maria Bethânia)

“A imaginação é mais importante que o conhecimento. O conhecimento é limitado, enquanto a imaginação abraça o mundo inteiro, estimulando o progresso, dando luz à evolução. Ela é, rigorosamente falando, um fator real na pesquisa científica.”

(Albert Einstein)

SUMÁRIO

RESUMO

O grupo *willistoni* de *Drosophila* serve como um modelo para investigar processos evolutivos na região Neotropical. O foco desta tese foi a intricada relação entre especiação, evolução do tamanho do genoma e elementos transponíveis (TEs) dentro deste grupo de espécies. Primeiramente, no Capítulo II foram abordadas as relações filogenéticas entre os seus subgrupos, com enfoque na parafilia do subgrupo *bocainensis*. Empregando marcadores moleculares, foi possível distinguir duas linhagens distintas dentro desse subgrupo, o qual ainda permanece parafilético. O Capítulo III explorou o papel da especiação na evolução do tamanho do genoma e da abundância de elementos repetitivos, com foco particular em TEs. Ao comparar o mobiloma entre espécies do subgrupo *willistoni*, sinais de mobilização recente foram detectados e possíveis impulsionadores da expansão desses genomas foram descritos, como os elementos da superfamília *TcMar-Tigger* de TEs. Esses resultados sugerem que processos de especiação em andamento podem contribuir para o aumento observado de elementos repetitivos e, consequentemente, do tamanho do genoma. No Capítulo IV, foram exploradas a variação do tamanho do genoma e a influência dos TEs na diversidade genômica entre linhagens de *D. willistoni*. Através de análises bioinformáticas, diferenças significativas no tamanho do genoma e na composição de elementos repetitivos entre linhagens foi detectada, com retrotransposons LTR desempenhando um papel crucial na diferença do tamanho desses genomas. Análises filogenéticas, ainda, sugeriram que duas subespécies ocorrem em simpatria na América do Sul. No Capítulo V, o papel dos TEs na mediação de rearranjos cromossômicos em *D. willistoni* foi elucidado. Através de uma busca e anotação de inversões cromossômicas induzidas por TEs, um grande número de inversões associadas a essas sequências foi sugerido. As principais famílias envolvidas incluem *Helitron*, *Gypsy* e o transposon *Galileo*, conhecido por induzir inversões em *D. buzzatii*. Esses achados sugerem que a atividade contínua de TEs é a principal impulsionadora da instabilidade genômica observada nessa espécie. Por último, no Capítulo VI, foi reconstruída a história evolutiva do transposon *Galileo*, extremamente abundante em *D. willistoni*. Foi encontrada uma distribuição ampla desse transposon nos genomas de Drosophilidae, além de eventos de transferência horizontal entre gêneros e grupos de espécies. Assim, esta tese avança a compreensão da evolução do tamanho do genoma, especiação e o papel dos TEs na formação da diversidade genômica dentro do grupo *willistoni*.

ABSTRACT

The *willistoni* group of *Drosophila* serves as a fundamental model for investigating evolutionary processes in the Neotropics. The focus of this thesis was the intricate relationship between speciation, genome size evolution, and transposable elements (TEs) within this group. First, in Chapter II, the phylogenetic relationships among its subgroups were addressed, with a focus on the paraphyly of the *bocainensis* subgroup. Through molecular markers, it was possible to distinguish two lineages within this subgroup, which still remains paraphyletic. Chapter III explored the role of speciation in shaping genome size evolution and the abundance of repetitive elements, with particular focus on TEs. By comparing the mobilome among species of the *willistoni* subgroup, signals of recent transposition were detected, and potential drivers of genome expansion were described, such as the *TcMar-Tigger* superfamily. These results suggest that ongoing speciation processes may contribute to the observed increase in repetitive elements and, consequently, genome size. In Chapter IV, genome size variation and the influence of TEs on genomic diversity among *D. willistoni* lineages were explored. Through bioinformatic analyses, significant differences in genome size and repetitive element composition among lineages were detected, with LTR retrotransposons playing a crucial role in genome size variation among populations of this species. Phylogenetic analyses suggest that two subspecies occur sympatrically in South America. In Chapter V, the role of TEs in mediating chromosomal rearrangements in *D. willistoni* was elucidated. Through a search and annotation of TE-induced inversions, a large number of inversions associated with these sequences were discovered. The main families include *Helitron*, *Gypsy*, and the *Galileo* transposon, known to induce inversions in *D. buzzatii*. These findings suggest that ongoing TE activity is the main driver of genomic instability observed in this species. Lastly, in Chapter VI, the evolutionary history of the *Galileo* transposon, extremely abundant in *D. willistoni*, was reconstructed. A widespread distribution of this transposon was found in Drosophilidae genomes, as well as horizontal transfer events between genera and species groups. Thus, the present thesis advances the understanding of genome size evolution, speciation, and the role of TEs in shaping genomic diversity within the *willistoni* group.

CAPÍTULO I

1. Introdução

1.1. Elementos transponíveis: o genoma em movimento

Durante o começo do século XX, as evidências obtidas sugeriam que os genes compreendiam porções fixas nos cromossomos (Morgan 1917). A visão de um genoma dinâmico começou a surgir somente na segunda metade do século, com as observações e estudos de Barbara McClintock (1902–1992). Entre as décadas de 1940 e 1950, a pesquisadora norte-americana descreveu padrões incomuns na herança da coloração em grãos de milho (*Zea mays* L.), relacionados a quebras no cromossomo 9 (McClintock 1939; Jones 2005). Suas observações levaram à descrição de um sistema onde dois *loci* associados (*Ac* e *Ds*) – os quais foram chamados de elementos controladores – mudavam de posição e causavam a quebra cromossômica observada (McClintock 1951, 1956). Seu trabalho permaneceu renegado por muitos anos por confrontar as ideias de Thomas Morgan (1917), já estabelecidas e amplamente aceitas na comunidade científica da época, acerca da estaticidade dos genes nos cromossomos. As descobertas de Barbara somente foram reconhecidas após a observação de elementos móveis na bactéria *Escherichia coli* na década de 1970 (Jones 2005), levando a pesquisadora a receber o Prêmio Nobel de Fisiologia e Medicina de 1983.

Chamados de elementos transponíveis (TEs – do Inglês, *transposable elements*), ou elementos de transposição, esses “genes saltadores” constituem sequências lineares de DNA que se deslocam nos genomas e que, em sua maioria, codificam enzimas próprias para realizarem esse processo – conhecido como mobilização (Bourque et al. 2018). Essas sequências variam tipicamente entre 100 a 10.000 pares de base de comprimento (Wells and Feschotte 2020), e são encontradas em diferentes proporções nos genomas de todos organismos pertencentes aos três domínios da vida (Figura 1). Até 60% do genoma é composto por TEs em espécies de eucariotos vertebrados (Sotero-Caio et al. 2017) – em *Homo sapiens*, por exemplo, essa fração chega aos 44% (Mills et al. 2007). Espécies vegetais, como o milho, e invertebrados, como *Drosophila melanogaster* Meigen, 1830, possuem, respectivamente, cerca de 85% e 20% de seu genoma composto por TEs (Kaminker et al. 2002; Schnable et al. 2009).

Essa grande proporção de TEs e quantidade de sequências encontradas culminou na necessidade de um sistema de classificação para catalogá-las e uma nomenclatura para

facilitar a comunicação entre a comunidade científica. O primeiro sistema, proposto por Finnegan (1989), divide os TEs em duas classes: aqueles que realizam sua transposição através da síntese de um RNA mensageiro (mRNA) intermediário, pertencentes à Classe I – retrotransposons; e aqueles que são transpostos diretamente em DNA, pertencentes à Classe II – transposons propriamente ditos (Figura 2). No primeiro caso, o TE codifica uma enzima transcriptase reversa (RT – do Inglês, *reverse transcriptase*) que faz a conversão do mRNA para DNA complementar (cDNA), o qual é inserido em outro local no genoma (Goodier 2016). A cópia original, portanto, permanece intacta e, agora, está duplicada; esse mecanismo de transposição é amplamente designado como "cópia e cola" (Göke and Ng, 2016). Por outro lado, os TEs da Classe II codificam uma enzima transposase que realiza a

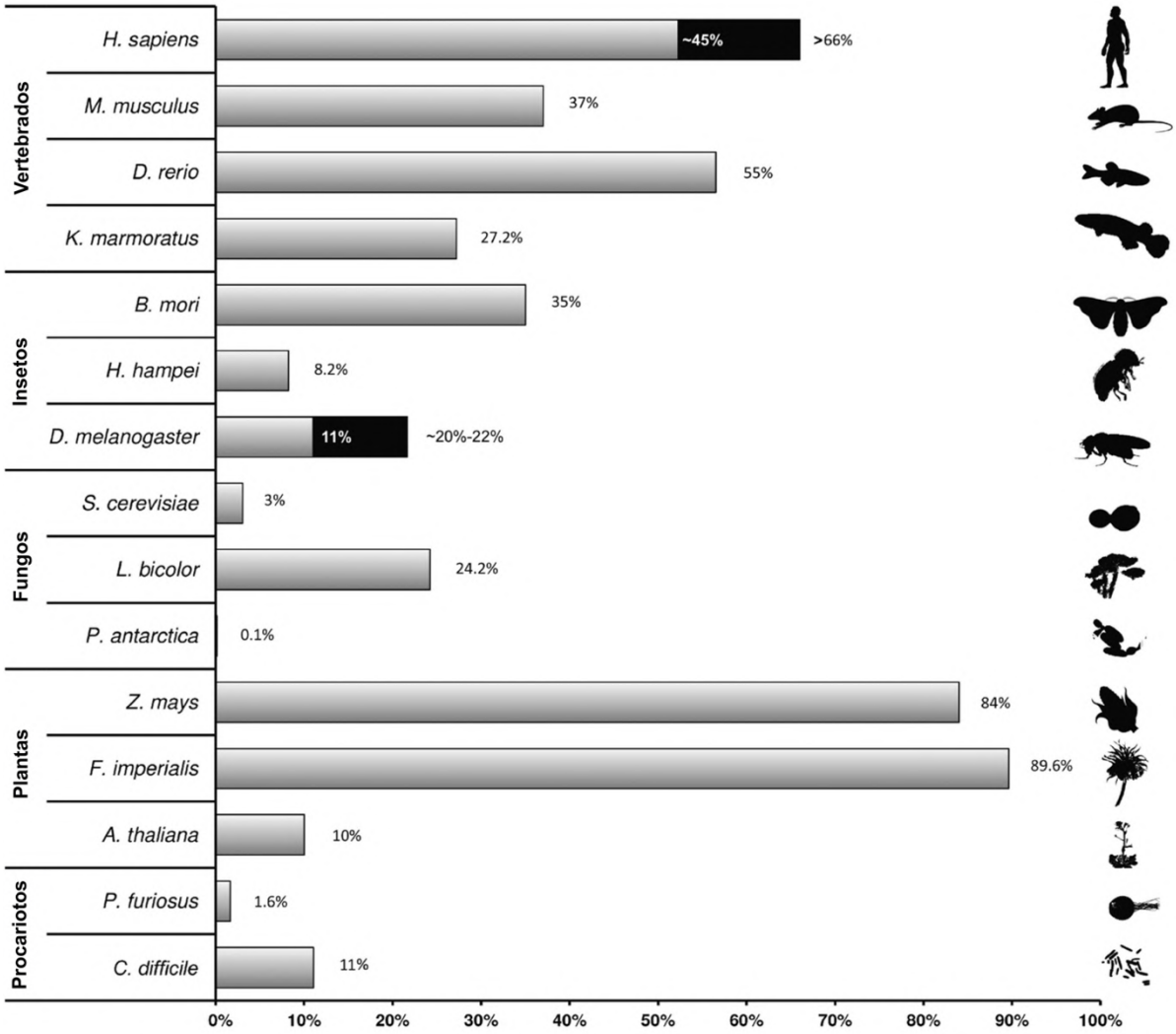


Figura 1. Porcentagem de elementos transponíveis em relação ao tamanho total do genoma em diferentes espécies eucarióticas e procarióticas. Barras na cor preta indicam variações entre diferentes estimativas. Adaptado de Guio & González (2019).

excisão do elemento e a posterior inserção em outro local – sendo esse processo chamado de “corta e cola” (Yusa et al. 2011).

Esse sistema foi confrontado com a descoberta de novos TEs que fugiam às regras, levando à criação de um sistema de classificação que concilia as duas classes e os novos critérios enzimáticos, proposto por Wicker et al. (2007). Nesse sistema, as duas classes são subdivididas em diferentes níveis, seguindo uma nomenclatura similar às taxonomias zoológica e botânica. As sequências de TEs continuam sendo agrupadas primeiramente em classes, porém, as subdivisões em ordens e superfamílias foram adicionadas, na maior parte dos casos com base em características estruturais (Piégu et al. 2015). Por último, seguindo critérios filogenéticos e de similaridade, a “regra 80-80-80” (ao menos 80% de identidade em no mínimo 80% da sequência a ser comparada, com no mínimo 80 pares de base) agrupa sequências em uma mesma família (Wicker et al. 2007).

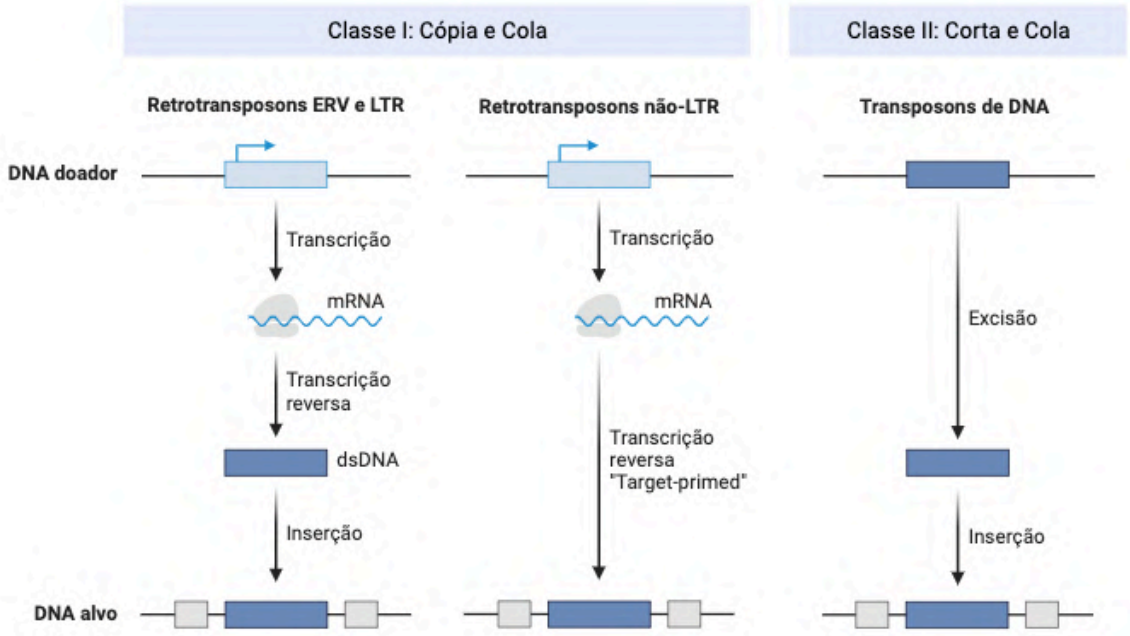


Figura 2. Mecanismos de transposição realizados por elementos transponíveis pertencentes às duas Classes. Nos elementos de Classe I, a transposição acontece com uma molécula de RNA mensageiro intermediária. Nos transposons de DNA, pertencentes à Classe II, a transposição é realizada diretamente em uma fita de DNA. Adaptado de Lanciano e Cristofari (2020) em BioRender.com.

Uma segunda classificação, complementar às demais, agrupa os TEs de acordo com a sua capacidade de mobilização (Wessler 2006). Elementos autônomos são aqueles que possuem seus genes intactos, sendo capazes de produzir, portanto, suas próprias enzimas para realizar a transposição (Wicker et al. 2007, 2021). Por outro lado, elementos não-

autônomos são cópias defectivas que não mais codificam suas próprias enzimas (Wicker et al. 2007, 2021). No entanto, TEs não-autônomos ainda podem ser mobilizados por proteínas expressas por TEs autônomos, evolutivamente próximos, localizados em outras partes do genoma – desde que o sítio de reconhecimento dessas proteínas ainda esteja intacto (Hayward & Gilbert, 2022).

1.2. Classe I – Retrotransposons

Há dois grupos principais de retrotransposons (Figura 3), caracterizados primeiramente pela topologia de seus genes e pelo mecanismo de transposição (Finnegan 2012), além da semelhança (ou não) a retrovírus (Wang and Han 2022). O primeiro corresponde à ordem LTR (do Inglês, *long terminal repeats*), englobando retroelementos que possuem nas suas extremidades terminações repetidas longas e diretas (LTRs). As LTRs possuem sequências de nucleotídeos repetitivos que desempenham um papel importante na regulação do elemento, porém não codificam proteínas. Elementos com LTR possuem uma estrutura básica com dois genes: *gag* e *pol* (Figura 3). O primeiro é assim denominado pois produz proteínas similares a capsídeos retrovirais, enquanto o último produz as demais enzimas necessárias para a transposição – integrase, RT e a RNAse H (Naville et al. 2016).

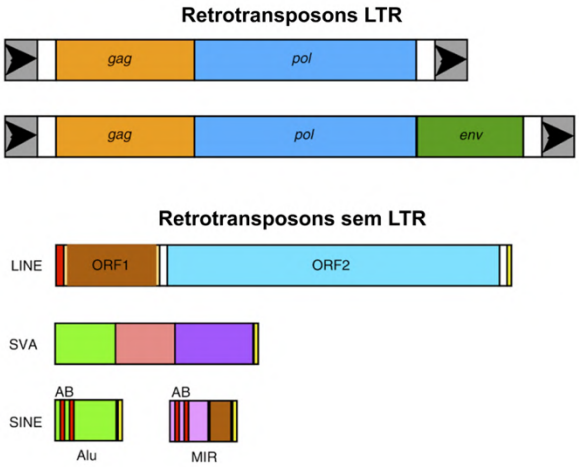


Figura 3. Estrutura de elementos típicos com e sem LTR. As LTRs constituem sequências idênticas e diretas (na mesma direção), localizadas em ambas extremidades de elementos com LTR. Nesses elementos, os genes *gag* e *pol* sintetizam, respectivamente, estruturas e enzimas necessárias para a transposição do retroelemento. Em elementos sem LTR, há a presença de duas fases de leitura aberta (ORFs), que codificam, respectivamente, a proteína *gag* e dois domínios: uma endonuclease e uma transcriptase reversa. Elementos SVA e SINE são elementos não-autônomos, por não possuírem ORFs. Adaptado de Finnegan (2012).

As ordens LINE (do Inglês, *long interspersed nuclear elements*) e SINE (do Inglês, *short interspersed nuclear elements*) correspondem aos "elementos sem LTR", pois apresentam apenas uma sequência de nucleotídeos poli-A em uma das suas terminações (Finnegan 2012). Esses TEs possuem duas fases de leitura aberta (ORFs – do Inglês, *open reading frame*) que codificam, respectivamente, a proteína *gag* e dois domínios: uma endonuclease (EM) e uma RT. Os elementos dessa ordem, diferente dos elementos com LTR, realizam a transposição através de um mecanismo chamado *target-primed reverse transcription* (Figura 2). Nesse mecanismo, a RT sintetiza a cópia de mRNA para DNA diretamente no sítio de integração localizado no cromossomo, gerando na extremidade 3' a característica cauda poli-A (Arkhipova and Yushenova, 2023). Os SINEs, por sua vez, são pequenos elementos não-autônomos que dependem da sintetização de enzimas por elementos LINE para sofrerem mobilização e transposição. Outras duas ordens, diferentes dos elementos com e sem LTR, pertencem à Classe I: os elementos DIRS (*Dictyostelium intermediate repeat sequence*) e PLE (*Penelope-like*) (Wicker et al. 2007). As principais características que os separam são mecanismos distintos de transposição; elementos DIRS não formam TSDs (do Inglês, *target site duplication*), enquanto elementos PLE apresentam somente regiões codificantes para RT e EN.

1.3. Classe II – Transposons de DNA

De uma forma geral, os transposons propriamente ditos realizam uma transposição dita não-replicativa, na qual o elemento é diretamente retirado do seu local e reinserido em outro. Essa classe é subdividida nas subclasses 1, que compreende as ordens TIR e Crypton, e a subclasse 2, que inclui as ordens Helitron e Maverick (Wicker et al. 2007). A ordem TIR (do Inglês, *terminal inverted repeats*) apresenta a estrutura mais simples entre as demais ordens de transposons (Figura 4). Esses TEs são constituídos por um gene que codifica a enzima transposase (TPase) e por repetições terminais invertidas (TIRs) em ambas extremidades. A TPase reconhece as TIRs e realiza a excisão do elemento, cortando as duas fitas de DNA, e reinsere o transposon em um sítio-alvo. O processo de inserção acaba gerando TSDs, as quais são únicas para cada família dessa subclasse. Dessa forma, as superfamílias são classificadas de acordo com a similaridade de suas TIRs ou TSDs. As ordens Helitron e Maverick, entretanto, realizam a quebra de apenas uma das fitas de DNA. Dessa forma, acabam realizando uma transposição replicativa.

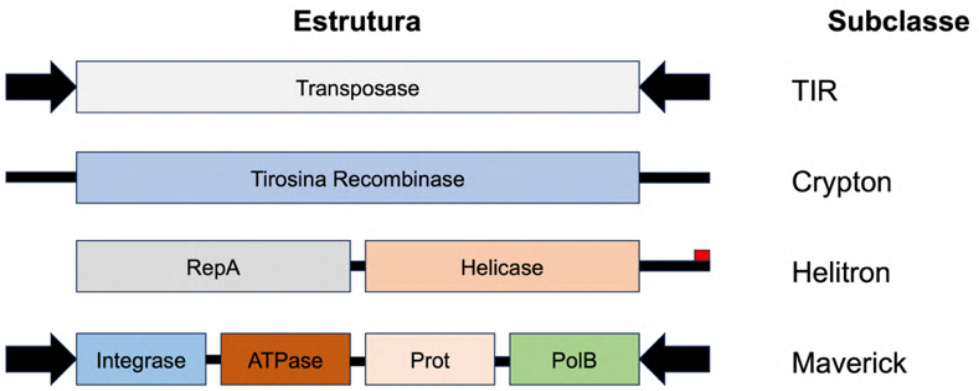


Figura 4. Organização e estrutura das quatro subclasses de transposons de DNA. Na subclasse TIR, a transposase é flanqueada pelas repetições terminais invertidas (TIRs), representadas pelas flechas. Helitrons possuem dois genes, a proteína de replicação A (RepA) e uma helicase; na extremidade 3', há uma estrutura *hairpin* (em vermelho). Elementos Maverick codificam as enzimas C-integrase, ATPase, cisteína protease (Prot) e DNA polimerase B (PolB). Adaptado de Wicker et al. (2007) e Piégu et al. (2015).

Há, ainda, um grupo heterogêneo de elementos não-autônomos de Classe II denominados MITEs (do Inglês, *Miniature Inverted-repeat Transposable Element*). Esses transposons podem ser mobilizados pela transposase de um outro transposon autônomo quando há similaridade em suas TIRs. Os MITEs, em particular, são pequenos elementos degenerados e derivados de transposons de DNA, com tamanho variando entre poucas dezenas a poucas centenas de pares de base (Fattash et al. 2013) entre outras características.

1.4. Ciclo de vida de TEs em genomas eucariotos

Assim como um parasita, os TEs possuem um ciclo de vida intimamente ligado com o genoma hospedeiro, onde a permanência e a garantia da passagem para a geração seguinte são fundamentais para a sua sobrevivência. O ciclo de vida de um TE (Figura 5) inicia com a sua invasão em um novo genoma, seja em uma mesma espécie via transferência vertical (parental), seja via transferência horizontal (HTT – do Inglês, *horizontal transposon transfer*) – quando o TE invade o genoma de uma outra espécie. Após a colonização há a proliferação do TE no genoma, o qual rapidamente aumenta seu número de cópias. Após sua amplificação no genoma hospedeiro, o TE vai proliferar dentro da população e acumular mutações, levando à sua diversificação. Após essa etapa, há dois cenários possíveis: o silenciamento, com ou sem equilíbrio nas taxas de transposição e seleção; ou a degradação do TE. No primeiro cenário, eventualmente o TE pode ser reativado e o ciclo recomeça. No último, há a possibilidade da domesticação do TE – criando novas sequências regulatórias;

ou a perda de função, quando os genes ou ORFs perdem a capacidade de codificar as suas enzimas. Entretanto, em todas as etapas é possível a ocorrência de eventos de HTT.

A transferência horizontal, particularmente, é um processo evolutivo que ocorre frequentemente em seres procariontes e envolve a passagem de genes, grupos de genes, plasmídeos e TEs entre espécies diferentes. Em eucariotos, entretanto, a transferência de material genético entre espécies distintas é menos comum. Os primeiros casos de HTT detectados em eucariotos foram recebidos com cautela, uma vez que desafiaram a visão clássica da hereditariedade vertical. Somente após o sequenciamento em larga escala do genoma completo de inúmeras espécies, os eventos de HTT tem sido corroborados e identificados com uma frequência muito maior do que previamente estimada (Schaack et al. 2010; Wallau et al. 2012).

A detecção de um possível evento de HTT ocorre a partir de três indícios: incongruência entre a filogenia dos TEs e aquela de suas espécies hospedeiras; alta similaridade entre TEs presentes em espécies filogeneticamente distantes; e uma distribuição irregular resultante da ausência do TE em algumas espécies. A inferência de um evento de HTT pode ser realizada ao se comparar o desvio do uso de códons (CUB – do Inglês, *codon usage bias*) entre espécies distintas (Wallau et al. 2016). Nesse caso, cada genoma apresentará um padrão único – incluindo os TEs nele inseridos. Assim, é possível comparar o CUB de um TE inserido em uma espécie com o CUB de outra. Em uma situação de HTT, o CUB será mais similar com o da outra espécie do que com o da espécie à qual pertence originalmente.

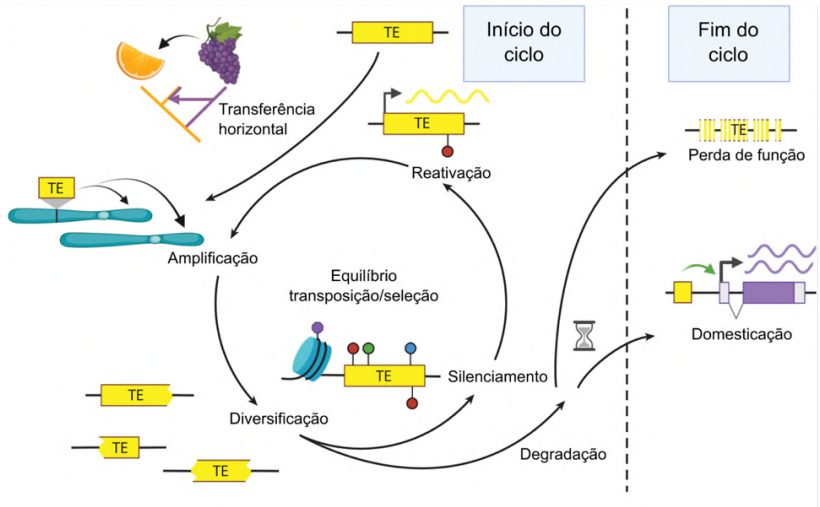


Figura 5. Ciclo de vida de um elemento transponível, em genomas eucariotos. Adaptado de Liu et al. (2022).

1.5. Uma fonte endógena de novidade evolutiva

Diante dos diversos mecanismos de mobilização, dos sítios de inserção e da própria presença de TEs espalhados pelo genoma, impactos tanto estruturais quanto fenotípicos são esperados como consequências da sua movimentação (Bourque et al. 2018). Um dos principais exemplos são as quebras cromossômicas e a mudança na coloração dos grãos de milho ocasionadas pelos transposons *Ac* e *Ds*, observados por Barbara McClintock. De fato, a inserção de um TE próximo a genes promove diversos desfechos possíveis. Nos genes ainda em funcionamento, a inserção de um TE próximo à região promotora pode resultar tanto na diminuição quanto no aumento da expressão (Zeng et al. 2018; Gebrie 2023). Além disso, a inserção dentro de regiões gênicas pode ocasionar o silenciamento total ao interromper a síntese de mRNA (Kobayashi et al. 2004). Por outro lado, TEs podem acabar reconstituindo genes já desativados, reintroduzindo funções genéticas previamente perdidas e contribuindo assim para a variabilidade e evolução do genoma (Bourque et al. 2018).

A dinâmica de transposição dos TEs modifica, também, a quantidade total do conteúdo de DNA genômico nos organismos (Becking et al. 2020; Zhang et al. 2020). A evolução do tamanho dos genomas, na verdade, é alvo de pesquisas desde a proposição do "paradoxo do valor-C" por Thomas (1971). Esse paradoxo se referia a falta de correlação (ou a correlação não encontrada) entre o valor-C (que corresponde ao tamanho total de um genoma, em sua versão haploide, em picogramas) e a "complexidade" dos organismos. Por exemplo, o valor-C no gênero protozoário *Amoeba* é cerca de 200 vezes maior que no gênero *Homo*; essa diferença, ainda, chega a 200 mil vezes entre organismos eucariotos (Gregory 2004). É necessário notar, contudo, que o significado de complexidade dos organismos na proposição desse paradoxo possui um viés antropocentrista ao colocar os seres humanos como espécie complexa em relação às demais. A busca de uma resposta para o paradoxo do valor-C levou pesquisadores a encontrar correlação entre o valor-C e diversas características, incluindo tamanho celular e corporal (Beaulieu et al. 2008), taxas metabólicas (Waltari and Edwards 2002), complexidade do desenvolvimento embrionário (Gregory 2002) e a história evolutiva das espécies (Jeffery et al. 2017). Nenhuma dessas, no entanto, é universal para todos os seres vivos.

Com a descoberta das sequências repetitivas e suas grandes proporções nos genomas, o paradoxo do valor-C foi, em parte, resolvido. O valor-C, na verdade, não representa a

quantidade total de genes em um genoma, uma vez que este é composto em sua maior parte por DNA não-codificante. Essas conclusões levaram à proposta do “enigma do valor-C” (Gregory 2001), um termo atualizado em relação ao paradoxo. Em geral, o enigma do valor-C traz como principal questão a variação na quantidade de DNA não-codificante encontrado nos genomas de eucariotos. Nesse caso, a atividade de elementos transponíveis (que fazem parte da fração não-codificante dos genomas) está diretamente relacionada ao aumento do tamanho total do genoma, uma vez que compreende fatores que ativamente adicionam novas sequências de DNA (Kidwell 2002; Gregory and Johnston 2008; Sessegolo et al. 2016).

Quanto à arquitetura cromossômica, os TEs são conhecidos por participarem na geração de rearranjos cromossômicos (Gray 2000). Inúmeros casos já foram descritos em diversas espécies, desde *Homo sapiens* (Balachandran et al. 2022) até *Drosophila* (Cáceres et al. 1999; Mérel et al. 2020). Os rearranjos cromossômicos surgem devido às quebras em um ou mais cromossomos, seguidas pela união subsequente das regiões afetadas de maneira reorganizada (Harewood et al. 2017). Quatro tipos principais de rearranjos estruturais podem ocorrer (Figura 6): (i) deleção, quando um segmento cromossômico é perdido e, junto com ele, toda a informação genética; (ii) duplicação, onde o segmento cromossômico (desde porções pequenas até o conjunto cromossômico completo) é duplicado; (iii) inversão, resultado da quebra das duas extremidades de um segmento cromossômico e do giro em 180º desta região; e (iv) translocação, ocorrendo após a troca de segmentos entre dois cromossomos não homólogos. As duas primeiras – também chamadas de rearranjos não-balanceados (Harewood et al. 2017) – acarretam mudanças drásticas na arquitetura genômica, pois resultam no ganho ou perda de material genético.

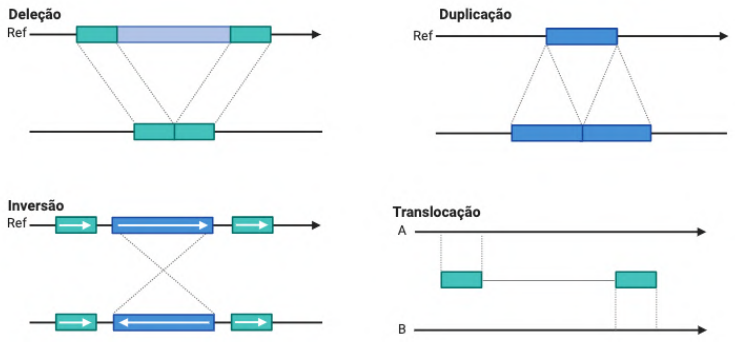


Figura 6. Esquema representando de uma forma geral os quatro principais tipos de rearranjos cromossômicos estruturais. Na deleção, o segmento cromossômico inteiro (em cinza) é perdido; na duplicação, o segmento cromossômico (em azul) é copiado e duplicado; na inversão, o segmento cromossômico (em azul) sofre um giro de 180º (conforme a orientação apresentada pelas flechas em branco) antes de ser reinserido no cromossomo; e na translocação, o segmento cromossômico (em verde) é trocado entre dois cromossomos diferentes (A e B). Produzido em Biorender.com.

As duas primeiras – também chamadas de rearranjos não-balanceados (Harewood et al. 2017) – acarretam mudanças drásticas na arquitetura genômica, pois resultam no ganho ou perda de material genético. As duas últimas costumam provocar apenas alterações estruturais, quando não ocorrem dentro de genes com funções essenciais. No gênero *Drosophila*, as famílias *Bel-Pao*, *Doc*, *FB*, *hobo*, *I*, *P* e *roo* foram encontradas associadas a recombinações ectópicas em populações de laboratório de *Drosophila melanogaster* (Lim and Simmons 1994); em populações naturais, o transposon *Galileo* foi responsável por gerar uma inversão em *D. buzzatii* (Cáceres et al. 1999).

1.6. *Galileo*, o transposon arquiteto

No cromossomo 2 de *D. buzzatii*, dois arranjos principais são encontrados: o ancestral, *2st*; e *2j*, o qual apresenta uma inversão em relação ao *2st* (Fontdevila et al. 1981). Ao sequenciar os pontos de quebra da inversão *2j*, Cáceres et al. (1999) encontraram grandes inserções que não existem no arranjo ancestral. Essas sequências nucleotídicas apresentavam indícios de pertencerem a um TE de Classe II: sítio-alvo duplicado (TSD) de 7 pares de base (bp); a presença de TIRs; e um padrão moderadamente repetitivo – mas distinto entre linhagens – nas hibridizações *in situ* em preparações de cromossomos politênicos. Além disso, a inversão e separação das sequências TSD produzidas durante a inserção do TE apontaram o mecanismo de recombinação ectópica (cruzamento entre duas sequências de DNA homólogas localizadas em posições cromossômicas não-alélicas) como responsável pela origem da inversão (Cáceres et al. 1999) – o que, mais tarde, foi confirmado por Delprat et al. (2009). Os autores, então, o nomearam *Galileo*.

Inicialmente, *Galileo* foi classificado como um transposon pertencente à família *foldback* devido a semelhanças estruturais (Cáceres et al. 2001; Casals et al. 2005) – como, por exemplo, as longas e internamente repetitivas TIRs, separadas por um domínio central, com tamanho e composição variados; e a capacidade de formar estruturas secundárias estáveis quando desnaturado – característica, essa, percebida pela dificuldade em amplificações via PCR. Outros dois transposons, *Kepler* e *Newton*, foram descritos durante a caracterização da inversão *2j* em outras linhagens de *D. buzzatii* (Cáceres et al. 2001). Algumas características indicaram sua estreita relação com *Galileo*, incluindo (i) a identidade nucleotídica média de 73%; os 40 bp terminais idênticos entre suas TIRs; ambos

os três transposons duplicam 7 bp durante a inserção; e as longas TIRs muito similares entre *Galileo* e *Newton*.

A primeira busca *in silico* por cópias de *Galileo* em outros genomas de *Drosophila* foi realizada por Marzo et al. (2008), facilitada pelo sequenciamento dos primeiros 12 genomas do gênero (*Drosophila* 12 Genomes Consortium, 2007). Foram encontradas sequências em seis espécies (*D. ananassae*, *D. mojavensis*, *D. persimilis*, *D. pseudoobscura*, *D. virilis* e *D. willistoni*), pertencentes aos dois grandes subgêneros de *Drosophila*. Além disso, as cópias encontradas mostraram ser muito degeneradas – todas sem a capacidade de codificar uma TPase funcional, sendo compostas ou somente por uma ou ambas TIRs, ou por TIRs e fragmentos do gene TPase (Figura 7). As cópias potencialmente inteiras descritas pelos autores, no entanto, foram montadas através de PCRs e apresentaram diversas mutações na fase de leitura (*frameshift*) e códons de terminação (*stop codons*).

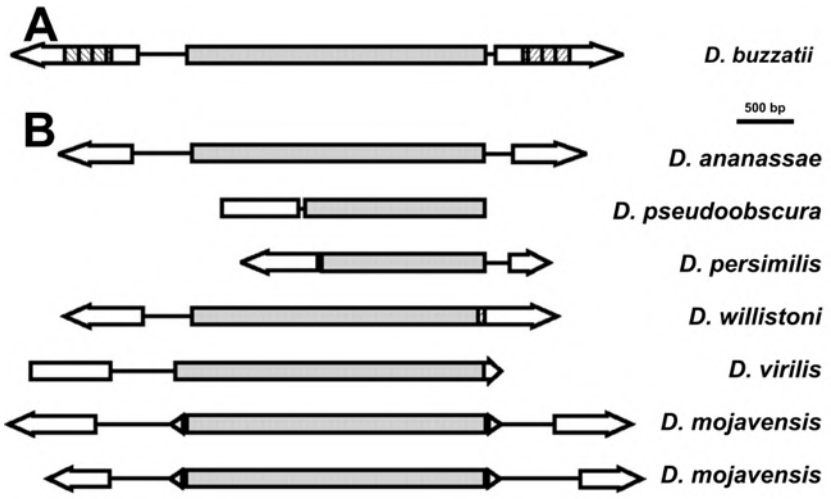


Figura 7. Estrutura de cópias do transposon *Galileo* encontradas por Marzo et al. (2008) em sete dos 12 genomas de espécies de *Drosophila* disponíveis até então. (A) Cópia potencialmente inteira de *Galileo* encontrada no genoma de *D. buzzatii*. (B) Cópias quase completas de *Galileo* em outras espécies. Setas representam as TIRs; retângulos cinza representam o gene da TPase; quadrados listrados representam as repetições diretas. Adaptado de Marzo et al. (2008).

Buscas utilizando a sequência de aminoácidos dessa cópia encontraram taxas altas de identidade entre a TPase de *Galileo* e as TPases dos transposons *1360* (também conhecido por *Hoppel*) e elemento *P* – incluindo o domínio THAP. Essas similaridades permitiram, portanto, a reclassificação da família *Galileo* como pertencente à superfamília *P*, junto às famílias *1360* e *P*; e a classificação dos transposons *Kepler* e *Newton* como subfamílias de *Galileo* no genoma de *D. buzzatii* (Marzo et al. 2008). Caracterizações mais detalhadas, além

das realizadas no genoma de *D. buzzatii* (Cáceres et al. 1999, 2001; Casals et al. 2003; Delprat et al. 2009), foram realizadas nos genomas de *D. mojavensis* (Marzo et al. 2013) e *D. willistoni* (Gonçalves et al. 2014). Em ambas espécies, um alto número de cópias foi encontrado – 170 e 191, respectivamente –, apresentando a mesma diversidade estrutural previamente observada por Marzo et al. (2008).

1.7. O grupo *willistoni* do gênero *Drosophila* (Diptera, Drosophilidae)

O gênero *Drosophila* pertence à Drosophilidae, uma família de insetos dípteros amplamente distribuídos em praticamente todos biomas terrestres, do nível do mar até grandes altitudes (Throckmorton 1975). Esses insetos se popularizaram no meio científico após a entrega do Prêmio Nobel de Medicina ou Fisiologia de 1933 à Thomas Morgan, cujo trabalho utilizou *Drosophila melanogaster* como organismo modelo e concluiu que os genes estavam localizados nos cromossomos (Morgan 1910). De fato, os drosofilídeos vêm sendo empregados como organismos modelos há mais de um século, desde o trabalho seminal de Castle et al. (1906). Dentre as diversas áreas do conhecimento, podemos destacar o seu uso em estudos genéticos, evolutivos, farmacológicos e ecológicos; este último, utilizando espécies como indicadoras de qualidade ambiental, por exemplo (Markow and O'Grady 2006; Valente-Gaiesky 2019).

A taxonomia dessas espécies possui uma intrincada história evolutiva subjacente, marcada por inúmeros eventos de especiação múltipla e subsequentes diversificações (Throckmorton 1975) (Figura 8). Análises moleculares (Yassin 2013) aliadas a caracteres morfológicos auxiliaram na resolução das relações filogenéticas no gênero *Drosophila*, reorganizando as radiações propostas por Throckmorton (1975) em diferentes subgêneros (Quadro 1).

Quadro 1. Reorganização proposta por Yassin (2013), com base em caracteres morfológicos e moleculares, para a taxonomia descrita por Throckmorton (1975) com base apenas em caracteres morfológicos. Adaptado de O'Grady & DeSalle (2018).

Throckmorton (1975)	**Yassin (2013)**
Radiação *virilis-repleta*	*Drosophila* (*Siphlodora*)
Radiação *immigrans-tripunctata*	*Drosophila* (*Drosophila*)
Drosophila (*Sophophora*)	*Drosophila* (*Sophophora*)
Drosophila (*Dorsilopha*)	*Drosophila* (*Dorsilopha*)

Até o momento, o gênero soma 1.661 espécies formalmente descritas; cerca de 34% da diversidade total de Drosophilidae (Bächli 2023). Essas espécies são, ainda, subdivididas em grupos e subgrupos de espécies, os quais agrupam espécies morfologicamente relacionadas (Grimaldi 1987; Grimaldi 1990). Como principal consequência, a árvore filogenética do gênero *Drosophila* é dividida não somente entre diversos subgêneros, mas também por outros gêneros que se diversificaram a partir de suas linhagens.

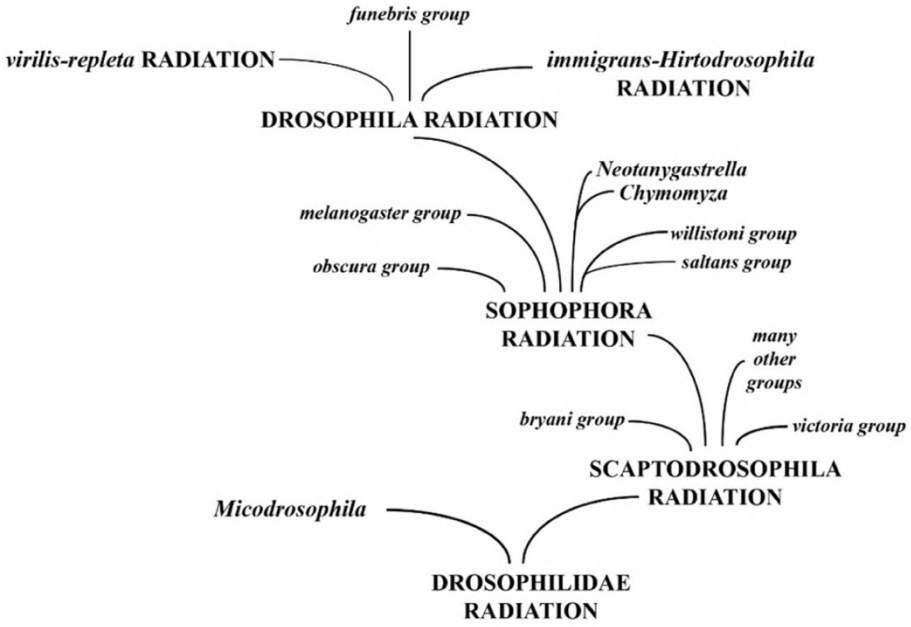


Figura 8. Representação clássica das relações filogenéticas e dos eventos de especiação múltipla na família Drosophilidae, com auxílio de caracteres morfológicos. Adaptado de Throckmorton (1975).

Em outras palavras, o gênero *Drosophila* apresenta, de acordo com sua taxonomia atual, relações essencialmente parafiléticas (Yassin 2013, Suvorov et al. 2022). O subgênero *Sophophora*, por exemplo, é uma linhagem formada por sete grupos de espécies e inclui o gênero *Lordiphosa*, irmão dos grupos *saltans* e *willistoni* (O'Grady and DeSalle, 2018; Suvorov et al. 2022). Curiosamente, esse gênero é originário das regiões tropicais e subtropicais asiáticas (O'Grady and DeSalle, 2018), enquanto os grupos *saltans* e *willistoni* compreendem o clado neotropical de *Sophophora* (O'Grady and Kidwell 2002).

O grupo *willistoni* abrange 24 espécies formalmente descritas, subdivididas nos subgrupos *alagitans* (seis espécies), *bocainensis* (12 espécies) e *willistoni* (seis espécies) (Quadro 1) (Bächli 2023). Estudos filogenéticos indicam sua monofilia (Robe et al. 2010; Zanini et al. 2018; Baião et al. 2023), no entanto utilizam dados morfológicos e/ou

moleculares das espécies do subgrupo *willistoni* e apenas quatro indivíduos do subgrupo *bocainensis*. Além disso, o subgrupo *bocainensis* é sempre recuperado como uma linhagem parafilética em relação ao subgrupo *willistoni*. A falta de estudos envolvendo os subgrupos *alagitans* e *bocainensis* se deve, principalmente, ao fato de que suas espécies são raramente amostradas com as técnicas usuais para captura de drosofilídeos (Salzano 1956) – de fato, há poucos registros para essas espécies, quando não apenas para o holótipo (Bächli 2023). O posicionamento desses subgrupos dentro do grupo *willistoni*, portanto, ainda permanece por ser esclarecido.

Quadro 2. Taxonomia e subdivisão das espécies do grupo *willistoni* de *Drosophila* nos subgrupos *alagitans*, *bocainensis* e *willistoni* (Bächli 2023).

Subgrupo *alagitans*	**Subgrupo *bocainensis***	**Subgrupo *willistoni***
D. alagitans	*D. abregolineata*	*D. equinoxialis*
D. capnoptera	*D. bocainensis*	*D. insularis*
D. megalagitans	*D. bocainoides*	*D. paulistorum*
D. neoalagitans	*D. capricorni*	*D. pavlovskiana*
D. neocapnoptera	*D. changuinolae*	*D. tropicalis*
D. pittieri	*D. fumipennis*	*D. willistoni*
	D. mangabeirai	
	D. nebulosa	
	D. parabocainensis	
	D. pseudobocainensis	
	D. subinfumata	
	D. sucinea	

As espécies que compõem o subgrupo *willistoni*, principalmente *D. willistoni*, são utilizadas há várias décadas como organismos modelo para a fauna neotropical, principalmente na área de genética evolutiva (revisão em Rohde and Valente, 2012). Elas formam um clado de espécies crípticas onde a distinção entre espécies ocorre, principalmente, pela análise de caracteres morfológicos internos, como as estruturas reprodutoras de indivíduos machos. Além disso, possuem distintos graus de isolamento reprodutivo e se apresentam em distintos níveis taxonômicos. No primeiro caso, a formação experimental de híbridos férteis entre *D. paulistorum* e *D. willistoni* foi reportada por Winge and Cordeiro (1963), bem como outros cruzamentos que geraram híbridos também férteis, inférteis ou que morreram em estágios iniciais do desenvolvimento (Winge 1965).

Quanto aos níveis taxonômicos, muitas espécies do subgrupo apresentam subespécies – como o caso de *D. paulistorum*, que compreende um complexo de seis subespécies que começaram a divergir nos últimos 2 milhões de anos, incluindo subespécies formadas há cerca de 300 mil anos (Zanini et al. 2018). Outro exemplo é *Drosophila willistoni*, formada por três subespécies – *D. w. quechua*, *D. w. willistoni* e *D. w. winge* (Mardiros et al. 2016), que estão distribuídas, respectivamente, a oeste da cordilheira dos Andes; na América Central, México, ilhas do Caribe e parte do Estado da Flórida, nos Estados Unidos; e no restante da América do Sul, até o norte da Argentina (Figura 9). De fato, *D. willistoni* é a espécie com a maior distribuição geográfica entre as do subgrupo (Zanini et al. 2015) e também é o drosofilídeo mais encontrado em florestas da América do Sul (Rohde and Valente 2012).

Figura 9. Distribuição aproximada das três subespécies de *Drosophila willistoni* na região neotropical, de acordo com Mardiros et al. (2016). Mapa produzido em Acme Mapper 2.2 (https://mapper.acme.com/).

Outra característica que tornou *D. willistoni* um modelo para estudos evolutivos são os altos níveis de polimorfismos encontrados em seus cromossomos: mais de 50 inversões descritas em três décadas; populações monomórficas estáveis, na verdade, nunca foram

encontradas em campo (revisão em Rohde and Valente 2012). O seu complemento cromossômico é constituído por dois pares de cromossomos metacêntricos (cromossomos II e X), um par acrocêntrico (cromossomo III) e um cromossomo Y submetacêntrico (Santos-Colares et al. 2003). O cromossomo III é resultado da fusão dos elementos E e F de Muller – sintênicos ao braço R do cromossomo 3 e ao cromossomo 4 de *Drosophila melanogaster* (Pita et al. 2014). Uma das principais hipóteses aponta os TEs como agentes primários causadores da instabilidade cromossômica observada nessa espécie.

Nesse sentido, o grupo *willistoni* se apresenta como um excelente modelo para a compreensão das dinâmicas de TEs e seus impactos nos genomas hospedeiros de espécies neotropicais. De fato, suas espécies possuem grande parte do genoma composto por sequências repetitivas, sendo *D. paulistorum* presente entre as espécies com maior porcentagem em Drosophilidae – com mais de 40% (Kim et al. 2021).

2. Objetivos

2.1. Objetivo geral: Contribuir para o entendimento sobre o papel de elementos transponíveis na evolução dos genomas de espécies incipientes no Neotrópico, utilizando as espécies do grupo *willistoni* de *Drosophila* como organismos modelo.

2.2. Objetivos específicos:

- Esclarecer as relações filogenéticas no grupo *willistoni*, com foco na parafilia do subgrupo *bocainensis* em relação ao subgrupo *willistoni* (Capítulo II);
- Fornecer um panorama acerca da influência do processo de especiação na evolução do tamanho de genoma e na sua fração repetitiva no subgrupo *willistoni* (Capítulo III);
- Investigar o papel dos elementos transponíveis na evolução do tamanho de genoma em nível intraespecífico em *D. willistoni* (Capítulo IV);
- Caracterizar as inversões cromossômicas possivelmente relacionadas com elementos transponíveis em uma linhagem de *D. willistoni* (Capítulo V);
- Propor uma hipótese para a história evolutiva do transposon *Galileo* nos genomas de Drosophilidae e investigar prováveis eventos de transferência horizontal (Capítulo VI).

CAPÍTULO II

Phylogenetic position of *Drosophila bocainensis* (Diptera, Drosophilidae) in the *willistoni* group and the paraphyletic status of the *bocainensis* subgroup

Henrique R.M. Antoniolli

Maríndia Deprá

Vera L.S. Valente

Manuscrito publicado no periódico *Canadian Journal of Zoology* (ISSN: 1480-3283)

doi: 10.1139/cjz-2023-0054

Canadian Journal of Zoology

Note

Phylogenetic position of *Drosophila bocainensis* (Diptera, Drosophilidae) in the *willistoni* group and the paraphyletic status of the *bocainensis* subgroup

Henrique R.M. Antoniolli[a,b], **Maríndia Deprá**[a,b], **and Vera L.S. Valente**[a,b]

[a]Laboratório de Drosophila, Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil; [b]Programa de Pós-Graduação em Genética e Biologia Molecular, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

Corresponding author: **Henrique R.M. Antoniolli** (email: antoniolli.henrique@hotmail.com)

Abstract

The *willistoni* group of *Drosophila* is subdivided into the *alagitans*, *bocainensis*, and *willistoni* subgroups, and has been an important model for studying evolutionary processes in the Neotropics for decades. Phylogenetic studies place the *bocainensis* subgroup as a grade in relation to the monophyletic *willistoni* subgroup, although these included molecular or morphological data for up to 4 species of the 12 species included in the first subgroup. Here, we characterized the first nucleotide sequences for three mitochondrial and five nuclear genes of *Drosophila bocainensis* Pavan & da Cunha, 1947 and employed these for addressing the paraphyly of this subgroup under a coalescent approach. Our results still recovered this paraphyletic relationship, placing *D. bocainensis*, *Drosophila capricorni* Dobzhansky & Pavan, 1943 and *Drosophila sucinea* Patterson & Mainland, 1944 in a basal clade, which diverged around 6.81 million years ago. The relationship of *Drosophila nebulosa* Sturtevant, 1916 and *Drosophila fumipennis* Duda, 1925 as a sister clade to the *willistoni* subgroup was recovered. The possible causes of such paraphyly are discussed.

Key words: coalescence, concordance factors, divergence times, *Drosophila bocainensis* Pavan & da Cunha, 1947 subgroup, Neotropics, paraphyly

Introduction

The *willistoni* group of *Drosophila* is one of the two Neotropical clades of the *Sophophora* subgenus (O'Grady and Kidwell 2002), and is currently subdivided into three subgroups: the *alagitans*, *bocainensis*, and *willistoni* subgroups (Bachli 2023). For several decades, the *willistoni* subgroup has been an important model for studying evolutionary processes in the Neotropical fauna, such as chromosomal rearrangements (see review in Rohde and Valente 2012) and reproductive isolation (review in Cordeiro and Winge 1995). However, as stated by Zanini et al. (2016), the *bocainensis* subgroup is often neglected or understudied in relation to the *willistoni* subgroup, yet the first plays a crucial role as an outgroup for studying the latter.

Many species belonging both to the *alagitans* and *bocainensis* subgroups, on the other hand, have not been largely sampled across its known geographical range, most of them presenting a few or only one record (Zanini et al. 2015). The latter also presents low frequencies in field samples, which may be a bias often related with the sampling efforts made with common banana baits for capturing drosophilids (Salzano 1956). This turns the description of new synapomorphies or DNA analyses into a hard task, and is the main impediment to deepening the knowledge about these species.

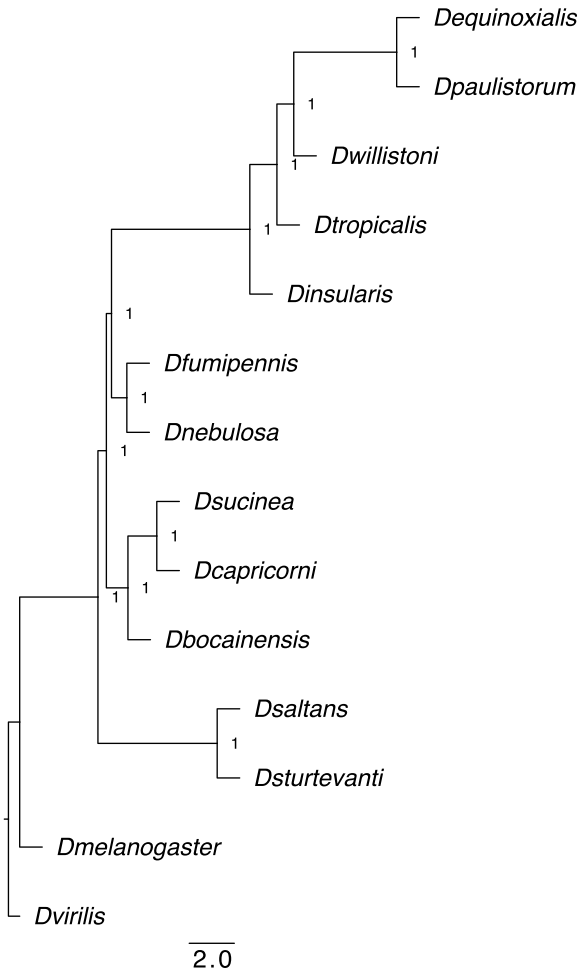
Currently, from the 12 species belonging to the *bocainensis* subgroup, only *Drosophila capricorni* Dobzhansky & Pavan, 1943, *Drosophila fumipennis* Duda, 1925, *Drosophila nebulosa* Sturtevant, 1916, and *Drosophila sucinea* Patterson & Mainland, 1944 have been studied at the molecular level. Phylogenetic studies performed so far recovered the *bocainensis* subgroup as a grade in relation to the *willistoni* subgroup (Gleason et al. 1998; Tarrío et al. 2000; Robe et al. 2010; Zanini et al. 2018; Finet et al. 2021). The position of each of the four species, however, floats between different phylogenies and employed molecular markers.

However, none of those studies have specifically addressed the paraphyly of the *bocainensis* subgroup, although some included the entire set of available data. Here, we assessed the phylogenetic position of *Drosophila bocainensis* Pavan & da Cunha, 1947 by characterizing the first nucleotide sequences for this species, and employed these to test the paraphyletic status of the *bocainensis* subgroup. We also report new divergence time estimates, especially for this subgroup, based on updated fossil calibrations (Suvorov et al. 2022).

Table 1. Specifications of annealing temperature and primer pairs used for each gene in this study.

Gene	Genetic compartment	Annealing temperature	Forward primer	Reverse primer	Reference
Cytochrome oxidase subunit I (*Cox1*)	Mitochondrial	57 °C	HCO1490	LCO2198	Folmer et al. (1994)
Cytochrome oxidase subunit II (*Cox2*)	Mitochondrial	55 °C	TL-J	TK-N	Simon et al. (1994)
Cytochrome *b* (*Cytb*)	Mitochondrial	55 °C	CB-J	TS1-N	Simon et al. (1994)
Alpha methyldopa hypersensitive (*Amd*)	Nuclear	Td 59–54 °C	AmdEx4F	Amd-bw	Tatarenkov et al. (2001); Robe et al. (2010)
Dopa decarboxylase (*Ddc*)	Nuclear	59 °C	BPF	BPR	Tatarenkov et al. (1999)
Male fertility factor (*kl-3*)	Nuclear	Td 60–52 °C	KL3-Y_F	KL3-Y_R	Zanini et al. (2018)
Hunchback (*Hb*)	Nuclear	Td 60–58 °C	Hb106F	Hb903R	Mota et al. (2008)
Period (*per*)	Nuclear	50 °C	per34	per28	Gleason and Powell (1997)

Fig. 1. Phylogram of the *willistoni* group of *Drosophila* inferred through a coalescent-based analysis in ASTRAL. Numbers next to nodes indicate the local posterior probability.



Materials and methods

DNA source, isolation, and sequencing

Flies of *D. bocainensis* were collected through banana traps in Erechim, southern Brazil (27°37′50.0″S, 52°14′11.0″W). Total genomic DNA was extracted from a single fly using the DNeasy Blood & Tissue® kit (QIAGEN, Hilden, Germany), following the manufacturer's protocol. Fragments of three mitochondrial and five nuclear genes were amplified with primer pairs described in Table 1. Polymerase chain reactions (PCR) were carried out using GoTaq® Hot Start Green Master Mix (Promega, Madison WI, USA), according to the manufacturer's protocol, with 0.2 μmol/L of each primer and 100–200 ng of DNA.

Amplifications for cytochrome oxidase subunit I (*Cox1*), cytochrome oxidase subunit I (*Cox2*), cytochrome *b* (*Cytb*), dopa decarboxylase (*Ddc*), and period (*per*) started with a step of denaturation at 94 °C for 5 min, followed by 35 cycles of denaturation at 94 °C for 45 s, annealing at specific temperatures (described in Table 1) for 45 s and extension at 72 °C for 1 min, and a final extension cycle at 72 °C for 5 min. For the alpha methyldopa hypersensitive (*Amd*), male fertility factor (*kl-3*), and hunchback (*Hb*) genes, we employed touchdown conditions, starting with a step of denaturation at 94 °C for 5 min, followed by 20 cycles of denaturation at 94 °C for 45 s, annealing at first temperature (Table 1) and extension at 72 °C for 1 min; 20 cycles with denaturation at 94 °C for 45 s, annealing at second temperature (Table 1) and extension at 72 °C for 1 min; and a final extension at 72 °C for 5 min. To check whether the amplification was successful, 5 μL of the 20 μL PCR product was submitted to electrophoresis in 0.8% agarose gel, stained with GelRed® (Biotium).

The obtained amplicons were purified with Exonuclease I (10 U/μL) and FastAP Thermosensitive Alkaline Phosphatase (1 U/μL) (Thermo Scientific Inc.) and directly sequenced in both 5′ and 3′ strands. Sequencing was performed by Macrogen Inc. (Seoul, South Korea) using BigDye technology and the same PCR primers. The electropherograms were assembled and inspected in the Gap4 software of the Staden Package (Staden 1996). The consensus sequences obtained were then checked in the BLASTn tool (NCBI website), and each polymorphic site was checked and corrected using the degeneracy table, if necessary.

Alignments and tests of nucleotide substitution saturation

Sequences of those eight genes plus the nuclear gene alcohol dehydrogenase (*Adh*) were retrieved from GenBank for five species belonging to the *willistoni* subgroup (*Drosophila*

Fig. 2. Plot of (A) mitochondrial genes tree versus (B) species tree based on the set of nuclear genes. Blue lines connect terminal taxa between trees, showing the different topologies for the *bocainensis* subgroup and for *Drosophila insularis* and *Drosophila tropicalis* in the *willistoni* subgroup.

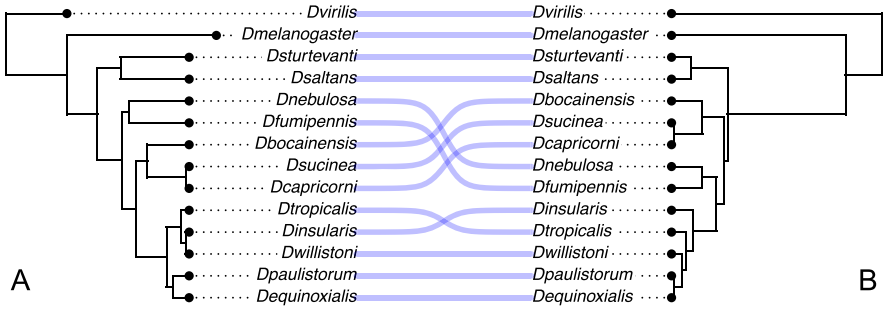


Table 2. Statistical results of tree topology tests comparing the ASTRAL species tree and an alternative tree, with a hypothetical monophyletic *bocainensis* subgroup, based on a concatenated alignment of all genes.

Tree	logL	deltaL	bp-RELL	*p*-KH	*p*-SH	*p*-WKH	*p*-WSH	c-ELW	*p*-AU
ASTRAL	−29 940.06872	0	0.971 +	0.968 +	1 +	0.968 +	0.968 +	0.967 +	0.979 +
Alternative	−29 954.21501	14.146	0.0292 −	0.0324 −	0.0324 −	0.0324 −	0.0324 −	0.0327 −	0.0209 −

Note: Plus signs denote the 95% confidence sets. Minus signs denote statistically significant exclusion. All tests performed 10 000 resamplings using the RELL method. deltaL, logL difference from the maximal logL in the set; bp-RELL, bootstrap proportion using RELL method (Kishino et al. 1990); *p*-KH, *p* value of one-sided Kishino–Hasegawa test(Kishino and Hasegawa 1989); *p*-SH, *p* value of Shimodaira–Hasegawa test (Shimodaira and Hasegawa 1999); *p*-WKH, *p* value of weighted Kishino–Hasegawa test; *p*-WSH, *p* value of weighted Shimodaira–Hasegawa test; c-ELW, expected likelihood weight (Strimmer and Rambaut 2002); p-AU, *p* value of approximately unbiased test (Shimodaira 2002).

paulistorum transitional (Zanini et al. 2018) representing the *D. paulistorum* species complex) and four species belonging to the *bocainensis* subgroup—the ingroup; *Drosophila sturtevanti* Duda, 1927 and *Drosophila saltans* Sturtevant, 1916 from the *saltans* group (sister clade to the *willistoni* group (Suvorov et al. 2022)) and *Drosophila melanogaster* Meigen, 1830 and *Drosophila virilis* Sturtevant, 1916—the outgroups (see accession numbers on Supplementary material 1—Table S1). Matrices were aligned in MAFFT online (Katoh et al. 2019) with automated strategy and adjusting direction according to the first sequence. The presence of paralogous copies of mitochondrial genes in the nucleus was evaluated by checking the presence of stop codons in mitochondrial markers. The alignments were then trimmed with TrimAl 1.4.1 (Capella-Gutiérrez et al. 2009) with the automated flag to remove poorly aligned regions. Potential saturation was assessed with the index of substitution saturation (I_{ss}) (Xia et al. 2003) implemented in DAMBE7 (Xia 2018).

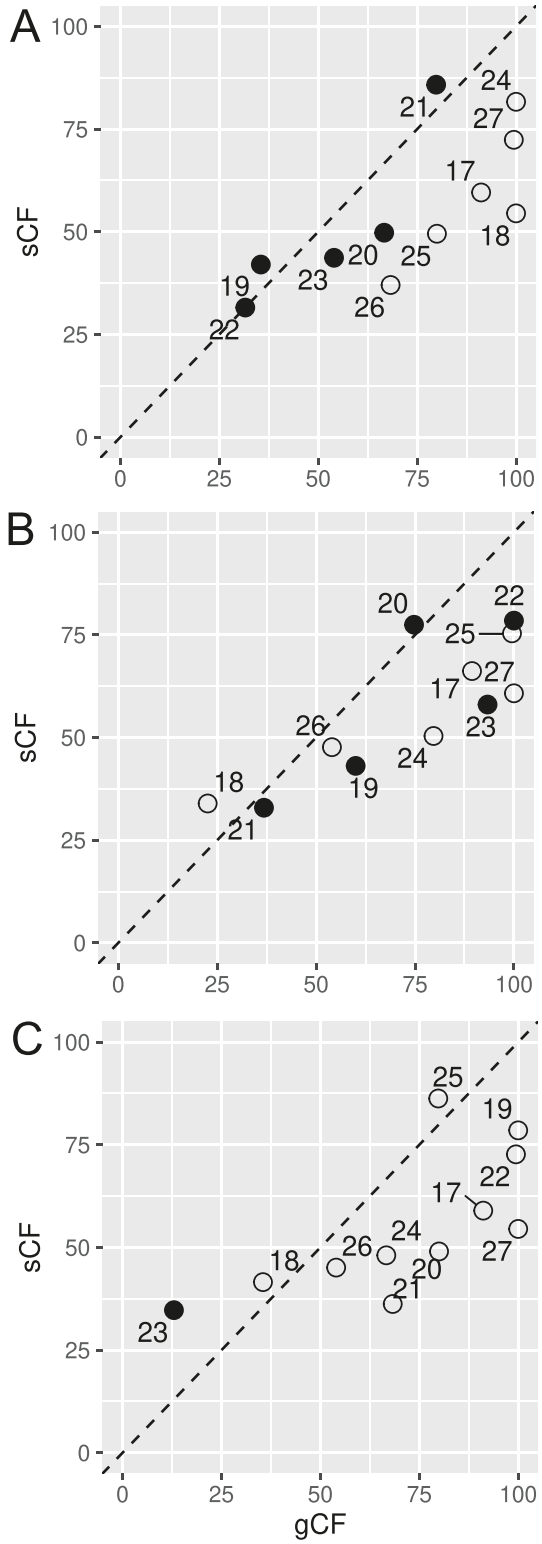
Inference of mitochondrial and species tree

The phylogenetic relationships were reconstructed through a Bayesian inference and coalescent-based analysis, as performed by Waichert et al. (2020). Phylogenetic trees were inferred in BEAST 1.10 (Suchard et al. 2018) individually for nuclear genes and simultaneously for mitochondrial genes, linking trees in the latter. The best nucleotide substitution model was estimated by ModelTest-NG (Darriba et al. 2020) based on Akaike Information Criterion (AIC) scores: GTR + I (*Cox2*, *Cytb*), GTR + G4 (*Amd*, *Ddc*, *Hb*, *kl-3*, *per*), and GTR + I + G4 (*Adh*, *Cox1*). An uncorrelated lognormal relaxed clock was applied, while the tree prior was set as Birth–Death process. Markov chain Monte Carlo (MCMC) searches were run for 5×10^7 generations, sampling every 5×10^3. Chain convergence was diagnosed in Tracer (Rambaut et al. 2018) and achieved if effective sample sizes values were higher than 200. Then, 100 trees were sampled for each nuclear gene and for the mitochondrial tree using LogCombiner, discarding the first 25% as burn-in. The resulting 700 trees were input into ASTRAL-III (Zhang et al. 2018) for reconstructing the species tree, whose root was placed in *D. virilis* when visualizing and editing in FigTree (https://github.com/rambaut/figtree).

Estimations of divergence times

Divergence times were estimated with *BEAST in BEAST 2 (Bouckaert et al. 2014) with the set of nuclear genes. Nucleotide substitution models, clock rates, and trees across the genes were left as unlinked. The substitution models were set as described in the previous section, while the tree prior was set as Calibrated Yule. The split between the *Drosophila* and *Sophophora* subgenera (represented by the split between *D. virilis* and *D. melanogaster*), according to a phylogenomic fossil-calibrated molecular clock (46.9 Myr, standard deviation of 1.5) (Suvorov et al. 2022), was set as calibration point. MCMC searches were run for 2×10^8 generations, sampling every 2×10^4. The maximum clade credibility tree was annotated with TreeAnnotator (Drummond and Rambaut 2007), with 25% of burn-in. The chronogram was plotted in R 4.2.1 (R Core Team 2023), with the packages *deeptime* (Gearty 2023), *ggplot2* (Wickham 2016), *ggtree* (Yu et al. 2017), and *treeio* (Wang et al. 2020).

Fig. 3. Plot of gene (gCF) and site (sCF) concordance factors for internal nodes on the (A) species tree based on the total set of genes, (B) species tree based on the set of nuclear genes, and (C) an alternative tree with a hypothetical monophyletic *bocainensis* subgroup. Filled circles correspond to the analyzed nodes, which are critical for the paraphyly of the *bocainensis* subgroup.



Tree topology and concordance factor tests

Comparative tests of topologies were performed in IQ-TREE 2 (Minh et al. 2020*a*) with the ASTRAL species tree and a manually built alternative tree, in which the *bocainensis* subgroup forms a monophyletic lineage. The alternative and the species trees had their branch lengths adjusted with treePL (Smith and O'Meara 2012), using as calibration point the same as described in the previous section. Both mitochondrial and nuclear genes were concatenated into a single supermatrix using FASconCAT-G (Kück and Longo 2014.). The Kishino–Hasegawa (Kishino and Hasegawa 1989), Shimodaira–Hasegawa (Shimodaira and Hasegawa 1999), and approximately unbiased (Shimodaira 2002) tests were performed with 10 000 resamplings using the RELL method.

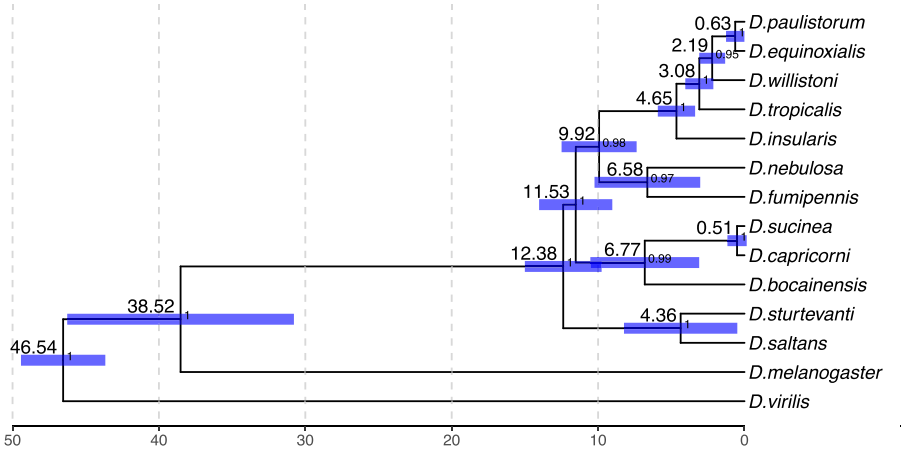
Additionally, the gene and site concordance factors (gCF and sCF, respectively) were estimated also in IQ-TREE 2. The gCF and sCF show the proportion of gene trees and sites in a concatenated matrix that are concordant with a given branch of a tree (Minh et al. 2020*b*). The first has a lower bound of 0% as its value is calculated from the full set of gene trees, while the latter has a lower bound of ~33%, given the three possible quartets for each node. These statistics were calculated for (i) the species tree reconstructed by ASTRAL with the entire set of genes; (ii) the species tree reconstructed by *BEAST with the nuclear genes; and (iii) the alternative tree with a hypothetical monophyletic *bocainensis* subgroup, with the entire set of genes.

Results and discussion

Here, the first nucleotide sequences of *D. bocainensis* were characterized and employed both to assess its phylogenetic position within the *willistoni* group and to address the paraphyletic status of the *bocainensis* subgroup. Divergence times were also updated for species belonging to this subgroup. The saturation tests were statistically significant; thus, little saturation was detected across all genes as the values of I_{ss} were lower than $I_{ss.c}$ (Supplementary material 2)—which is congruent with the results reported by Zanini et al. (2018). Therefore, the inclusion of *D. bocainensis* and the outgroups in the matrices kept the phylogenetic information obtained in that previous study.

The species tree topology (Fig. 1) recovered for the *willistoni* subgroup is consistent with previously established relationships using concatenation methods (Robe et al. 2010; Zanini et al. 2018; Finet et al. 2021). The first species to branch off is *Drosophila insularis* Dobzhansky, 1957, followed by *Drosophila tropicalis* Burla & da Cunha, 1949, *Drosophila willistoni* Sturtevant, 1916, *Drosophila equinoxialis* Dobzhansky, 1946, and *Drosophila paulistorum* Dobzhansky & Pavan, 1949 (local posterior probability (PP) = 1.0). The inclusion of *D. bocainensis*, however, was not able to recover the monophyly of its subgroup, which remained as a grade: *D. fumipennis* and *D. nebulosa* in a clade sister to the *willistoni* subgroup (local PP = 1.0), another clade with *D. bocainensis* as the first to branch off (local PP = 1.0), and *D. capricorni* and *D. sucinea* as sister species (local PP = 1.0).

Fig. 4. Chronogram recovered by *BEAST representing the divergence times across the phylogeny of the *willistoni* group of *Drosophila*. Numbers above nodes indicate the mean age (in million years) of each split, whose 95% confidence limits are reflected by the respective node bars; numbers on the side of each node correspond to its posterior probability.



Nonetheless, the mitochondrial tree (Supplementary material 3—Fig. S1) recovered a different topology for the *bocainensis* subgroup (Fig. 2A). In this case, the lineage containing *D. bocainensis* was placed as sister to the *willistoni* subgroup (PP = 1.0) and the clade *D. fumipennis* and *D. nebulosa* (PP = 0.54) as sister to them (PP = 1.0). Relationships within the *willistoni* subgroup were also quite different, splitting (PP = 1.0) this subgroup into two clades: *D. paulistorum* and *D. equinoxialis* (PP = 0.99); and *D. tropicalis* (as the first to branch off), *D. insularis*, and *D. willistoni* (PP = 1.0). All topology tests rejected the alternative tree in favor of the topology reconstructed by ASTRAL-III (Table 2), supporting the paraphyletic relationships found in the *bocainensis* subgroup. In these analyses, a *p* value lower than 0.05 indicates the rejection of a given topology. Gene and site concordance factors added further support to the species tree, and their results are focused on the nodes 19 (the clade *willistoni* group), 20 (the clade *D. bocainensis* + *D. capricorni* + *D. sucinea*), 22 (the clade *willistoni* subgroup + *D. fumipennis* + *D. nebulosa*), and 23 (*D. fumipennis* + *D. nebulosa*) in the nuclear and species tree and 23 (the hypothetical monophyletic *bocainensis* subgroup) in the alternative tree, as these nodes are critical for assessing the paraphyly of the *bocainensis* subgroup. In all comparisons, gCF and sCF were correlated (Fig. 4).

First, the species tree with the entire set of genes (Fig. 4A) showed lower scores of gCF and sCF (for most nodes) than the species tree with nuclear genes only (Fig. 4B) for all the evaluated nodes (Supplementary materials 4 and 5—Tables S2 and S3). The exceptions are the node 20, which had a support of 49.77% of sCF in the total set compared with 49.03% of sCF in the nuclear set, and the node 23, with a gCF of 54% and an sCF of 43.66% in the first versus 13% of gCF and 34.74% of sCF in the latter. However, the node 20 had an increase in gCF—from 66.67% in the total set to 80% in the nuclear set—providing extra support for the lineage of *D. bocainensis*. The node 23 in the alternative topology (Fig. 4C) had also a low sCF (34.74%), but the lowest gCF across all comparisons (13%); i.e., only 13% of the 700 trees support monophyly for the *bocainensis* subgroup.

Mitochondrial markers have previously recovered distinct topologies in comparison with nuclear genes for the *willistoni* group (Gleason et al. 1998; O'Grady and Kidwell 2002; Robe et al. 2010; Zanini et al. 2018). The interchange between the lineage of *D. bocainensis* and that of *D. fumipennis* as sister group to the *willistoni* subgroup is probably an outcome of the multiple introgressions that may have occurred in its evolutionary history (Baião et al. 2023). Chromosome banding patterns found by Pita et al. (2014) also placed *D. nebulosa* as closely related to the *willistoni* subgroup and suggested that *D. capricorni* and *D. fumipennis* belong to different lineages—as seen in the nuclear and mitochondrial gene trees (Fig. 2). Other phylogenies (Finet et al. 2021) also agreed with the position of *D. nebulosa*, which supports that the lineage containing *D. nebulosa* and *D. fumipennis* is the most likely sister group to the *willistoni* subgroup.

The topology of the chronogram (Fig. 3), which was reconstructed with the set of nuclear genes under the multi-species coalescent model, is identical to the species tree (PP > 0.95). The divergence times in the *willistoni* group were first estimated by Robe et al. (2010), based on the evolutionary rate of *Adh* estimated by Russo et al. (1995), followed by the estimations of Zanini et al. (2018) based on fossil calibrations. Recently, Suvorov et al. (2022) employed several schemes of fossil and molecular calibrations to estimate the divergence times across Drosophilidae, in a phylogenomic framework, which allowed us to perform new estimations. Our results showed that the diversification of the clade containing *D. bocainensis*, *D. capricorni*, and *D. sucinea* began around 6.81 (95% confidence interval (CI): 10.16–2.73) million years ago (Mya). The split between *D. capricorni* and *D. sucinea* seems to be more recent (0.51 Mya; 95% CI: 0.0002–1.32) than the estimation of Zanini et al. (2018) (3.55 Mya; 95% CI: 6.45–1.31); the other divergence times also differed, but fell into their 95% CI. This is the case of the clade containing *D. fumipennis* + *D. nebulosa*

and the *willistoni* subgroup (PP = 0.98), whose diversification occurred around 9.92 Mya (95% CI: 7.44–12.56). Similar to the lineage of *D. bocainensis*, the split between *D. fumipennis* and *D. nebulosa* occurred 6.58 Mya (95% CI: 2.73–9.96). These results reinforce that the *willistoni* group experienced recent speciation events (Zanini et al. 2018).

Correctly establishing the phylogenetic relationships between species is crucial for understanding their evolutionary history. This becomes even more critical when studying rare species, such as those belonging to the *bocainensis* subgroup (Salzano 1956). Rare species are essential for ecosystem functioning in terms of functional diversity or functional redundancy (Jain et al. 2013), and have impact on high trophic levels and on their community structure (Bracken and Low 2012). Overall, the data currently available are not sufficient to formally split the *bocainensis* subgroup into two, especially due to the low number of sampled species. However, the present evidence—the high support on the species tree (local PP = 1.0) and on the chronogram (PP > 0.98), the rejection of a monophyletic topology, and the concordance factors—at least indicates that there are two monophyletic lineages within the *bocainensis* subgroup. This should be taken into consideration when studying these species themselves or employing them as outgroups for the *willistoni* subgroup. A thorough systematic review comprising the entire set of species and employing more molecular markers should clarify whether the two lineages of the *bocainensis* subgroup form a clade or a grade.

Acknowledgements

The authors thank Dr Lizandra Jaqueline Robe and Dr Sebastián Pita for helpful comments during the development of this study.

Article information

History dates

Received: 14 March 2023
Accepted: 29 July 2023
Accepted manuscript online: 31 August 2023
Version of record online: 16 October 2023

Copyright

Data availability

The data produced in this study are deposited in GenBank (NCBI), under accession nos. OQ351288-89 and OQ357798-7806.

Author information

Author ORCIDs

Henrique R.M. Antoniolli https://orcid.org/0000-0002-6747-3641
Maríndia Deprá https://orcid.org/0000-0003-4568-1869
Vera L.S. Valente https://orcid.org/0000-0001-8661-6284

Author contributions

Data curation: HRMA
Formal analysis: HRMA
Investigation: HRMA, MD, VLSV
Supervision: MD, VLSV
Writing – original draft: HRMA
Writing – review & editing: MD, VLSV

Competing interests

The authors declare there are no competing interests.

Funding information

This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under the scholarship no. 141319/2020-8 and research productivity grant no. 312781/2018-0. This study was also financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES).

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/cjz-2023-0054.

References

Bächli, G. 2023. TaxoDros. Available from https://www.taxodros.uzh.ch/ [accessed 23 January 2023].

Baião, G.C., Schneider, D.I., Miller, W.J., and Klasson, L. 2023. Multiple introgressions shape mitochondrial evolutionary history in *Drosophila paulistorum* and the *Drosophila willistoni* group. Mol. Phylogenet. Evol. **180**: 107683. doi:10.1016/j.ympev.2022.107683. PMID: 36574824.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C-H., Xie, D., et al. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. **10**: e1003537. doi:10.1371/journal.pcbi.1003537. PMID: 24722319.

Bracken, M.E.S., and Low, N.H.N. 2012. Realistic losses of rare species disproportionately impact higher trophic levels. Ecol. Lett. **15**(5): 461–467. doi:10.1111/j.1461-0248.2012.01758.x. PMID: 22381064.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, **25**(15): 1972–1973. doi:10.1093/bioinformatics/btp348. PMID: 19505945.

Cordeiro, A.R., and Winge, H. 1995. Levels of evolutionary divergence of *Drosophila willistoni* sibling species. *In* Genetics of natural populations of Theodosius Dobzhansky. *Edited by* L. Levine. Columbia University Press, New York. pp. 241–261.

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol. Biol. Evol. **37**(1): 291–294. doi:10.1093/molbev/msz189. PMID: 31432070.

Drummond, A.J., and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**: 214. doi:10.1186/1471-2148-7-214. PMID: 17996036.

Finet, C., Kassner, V.A., Carvalho, A.B., Chung, H., Day, J.P., Day, S., et al. 2021. DrosoPhyla: resources for Drosophilid phylogeny and systematics. Genome Biol. Evol. **13**(8): evab179. doi:10.1093/gbe/evab179. PMID: 34343293.

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. **3**(5): 294–299. PMID: 7881515.

Gearty, W. 2023. deeptime: plotting tools for anyone working in deep time. R package version 1.0.0.Available from https://CRAN.R-project.org/package=deeptime [accessed 10 March 2023].

Gleason, J.M., and Powell, J.R. 1997. Interspecific and intraspecific comparisons of the *period* locus in the *Drosophila willistoni* sibling species.

Mol. Biol. Evol. **14**(7): 741–753. doi:10.1093/oxfordjournals.molbev.a025814. PMID: 9214747.

Gleason, M.J., Griffith, E.C., and Powell, J.R. 1998. A molecular phylogeny of the *Drosophila willistoni* group: conflicts between species concepts? Evolution, **52**(4): 1093–1103. doi:10.2307/2411239. PMID: 28565231.

Jain, M., Flynn, D.F.B., Prager, C.M., Hart, G.M., DeVan, C.M., Ahrestani, F.S., et al. 2013. The importance of rare species: a trait-based assessment of rare species contributions to functional diversity and possible ecosystem function in tall-grass prairies. Ecol. Evol. **4**(1): 104–112. doi:10.1002/ece3.915. PMID: 24455165.

Katoh, K., Rozewicki, J., and Yamada, K.D. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings Bioinf. **20**(4): 1160–1166. doi:10.1093/bib/bbx108. PMID: 28968734.

Kishino, H., Miyata, T., and Hasegawa, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol. Evol. **31**: 151–160. doi:10.1007/BF02109483.

Kishino, H., and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29**: 170–179. doi:10.1007/BF02100115. PMID: 2509717.

Kück, P., and Longo, G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. **11**: 81. doi:10.1186/s12983-014-0081-x. PMID: 25426157.

Minh, B.Q., Hahn, M.W., and Lanfear, R. 2020b. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. **37**(9): 2727–2733. doi:10.1093/molbev/msaa106. PMID: 32365179.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. 2020a. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37**(5): 1530–1534. doi:10.1093/molbev/msaa015. PMID: 32011700.

Mota, N.R., Robe, L.J., Valente, V.L.S., Budnik, M., and Loreto, E.L.S. 2008. Phylogeny of the *Drosophila mesophragmatica* group (Diptera, Drosophilidae): an example of Andean evolution. Zool. Sci. **25**: 526–532. doi:10.2108/zsj.25.526. PMID: 18558806.

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**(3): 442–453. doi:10.1006/mpev.2001.1053. PMID: 11884169.

Pita, S., Panzera, Y., Valente, V.L.S, Melo, Z.G.S., Garcia, C., Garcia, A.C.L., et al. 2014. Cytogenetic mapping of the Muller F element genes in *Drosophila willistoni* group. Genetica, **142**: 397–403. doi:10.1007/s10709-014-9784-3. PMID: 25134938.

R Core Team. 2023. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/ [accessed 23 January 2023].

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst. Biol. **67**(5): 901–904. doi:10.1093/sysbio/syy032. PMID: 29718447.

Robe, L.J., Cordeiro, J., Loreto, E.L.S., and Valente, V.L.S. 2010. Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila willistoni* subgroup (Diptera: Drosophilidae). Genetica, **138**: 601–617. doi:10.1007/s10709-009-9432-5. PMID: 20049511.

Rohde, C., and Valente, V.L.S. 2012. Three decades of studies on chromosomal polymorphism of *Drosophila willistoni* and description of fifty different rearrangements. Genet. Mol. Biol. **35**(4): 966–979. doi:10.1590/S1415-47572012000600012. PMID: 23411997.

Russo, C.A.M., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence times of Drosophilid species. Mol. Biol. Evol. **12**: 391–404. doi:10.1093/oxfordjournals.molbev.a040214. PMID: 7739381.

Salzano, F.M. 1956. Chromosomal polymorphism and sexual isolation in sibling species of the *bocainensis* subgroup of *Drosophila*. Evolution, **10**(3): 288–297. doi:10.1111/j.1558-5646.1956.tb02853.x.

Shimodaira, H., and Hasegawa, M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. **16**: 1114. doi:10.1093/oxfordjournals.molbev.a026201.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51**(3): 492–508. doi:10.1080/10635150290069913. PMID: 12079646.

Shimodaira, H., and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**(8): 1114. doi:10.1093/oxfordjournals.molbev.a026201.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., and Flook, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. **87**(6): 651–701. doi:10.1093/aesa/87.6.651.

Smith, S.A., and O'Meara, B.C. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics, **28**(20): 2689–2690. doi:10.1093/bioinformatics/bts492. PMID: 22908216.

Staden, R. 1996. The staden sequence analysis package. Mol. Biotechnol., **5**: 233–241. doi:10.1007/BF02900361. PMID: 8837029.

Strimmer, K., and Rambaut, A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. R Soc. Lond. B **269**: 137–142. doi:10.1098/rspb.2001.1862.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. **4**(1): vey016. doi:10.1093/ve/vey016. PMID: 29942656.

Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., et al. 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr. Biol. **32**: 111–123. doi:10.1016/j.cub.2021.10.052. PMID: 34788634.

Tarrío, R., Rodríguez-Trelles, F., and Ayala, F.J. 2000. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. Mol. Phylogenet. Evol. **16**(3): 344–349. doi:10.1006/mpev.2000.0813. PMID: 10991788.

Tatarenkov, A., Kwiatowski, J., Sharecky, D., Barrio, E., and Ayala, F.J. 1999. On the evolution of DOPA decarboxylase (*Ddc*) and *Drosophila* systematics. J. Mol. Evol. **48**: 445–462. doi:10.1007/PL00006489. PMID: 10079283.

Tatarenkov, A., Zurovcova, M., and Ayala, F.J. 2001. *Ddc* and *Amd* sequences resolve phylogenetic relationships of *Drosophila*. Mol. Phylogenet. Evol. **20**: 321–325. doi:10.1006/mpev.2001.0967. PMID: 11476641.

Waichert, C., Wilson, J.S., Pitts, J.P., and von Dohlen, C.D. 2020. Phylogenetic species delimitation for the widespread spider wasp *ageniella accepta* (Hymenoptera: Pompilidae), with new synonyms. Insect Syst. Evol. **51**: 532–549. doi:10.1163/1876312X-00002207.

Wang, L.G., Lam, T.T.Y., Xu, S., Dai, Z., Zhou, L., Feng, T., et al. 2020. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol. Biol. Evol. **37**(2): 599–603. doi:10.1093/molbev/msz240. PMID: 31633786.

Wickham, H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York. Available from https://ggplot2.tidyverse.org.

Xia, X. 2018. DAMBE7: new and improved tools for data analysis in molecular biology and evolution. Mol. Biol. Evol. **35**: 1550–1552. doi:10.1093/molbev/msy073. PMID: 29669107.

Xia, X., Xie, Z., Salemi, M., Chen, L., and Wang, Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. **26**: 1–7. doi:10.1016/S1055-7903(02)00326-3. PMID: 12470932.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. **8**(1): 28–36. doi:10.1111/2041-210X.12628.

Zanini, R., Deprá, M., and Valente, V.L.S. 2015. On the geographic distribution of the *Drosophila willistoni* group (Diptera, Drosophilidae)—updated distribution of *alagitans* and *bocainensis* subgroups. Drosophila Inf. Serv. **98**: 25–27.

Zanini, R., Deprá, M., and Valente, V.L.S. 2016. Ultrastructural characterization of the pre adult stages of *Drosophila willistoni* species group (Diptera, Drosophilidae). Trends Entomol. **12**: 43–50.

Zanini, R., Müller, M.J., Vieira, G.C., Valiati, V.H., Deprá, M., and Valente, V.L.S. 2018. Combining morphology and molecular data to improve *Drosophila paulistorum* (Diptera, Drosophilidae) taxonomic status. Fly, **12**(2): 81–94. doi:10.1080/19336934.2018.1429859. PMID: 29355090.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. **19**(153): 15–62. doi:10.1186/s12859-018-2129-y. PMID: 29343218.

Supplementary material

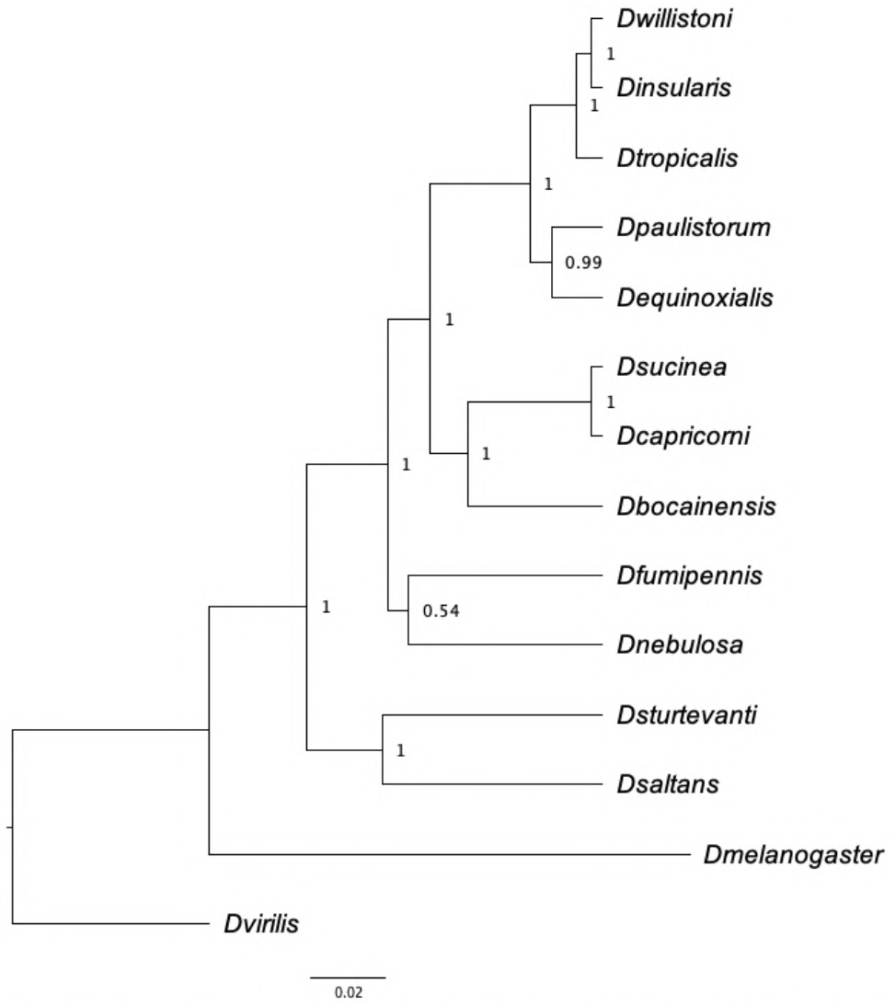


Figure S1. Phylogram of the *willistoni* group of *Drosophila* inferred by mitochondrial markers through Bayesian Inference in BEAST 1. Numbers next to nodes indicate the posterior probability (PP).

Table S1. Taxonomy of each species included in the phylogenetic analysis and GenBank accession numbers of each nucleotide sequence.

Group	Subgroup	Species	Subspecies	Cox1	Cox2	Cytb	Amd	Adh	Ddc	kl3-Y	Hb	per
willistoni	*willistoni*	*D. equinoxialis*		EU493638.1	EU493767.1	EU494151.1	FJ664506.1	MG010087.1	MG010058.1	MG010072.1	EU532101.1	U51074.1
		D. insularis		MG010091.1	MG010106.1	MG010109.1	FJ664507.1	EU532120.1	MG010059.1	MG010073.1	EU532098.1	U51088.1
		D. paulistorum	*transitional*	MG010099.1	EU532093.1	MG010117.1	FJ664501.1	EU532130.1	MG010067.1	MG010081.1	EU532104.1	U51085.1
		D. tropicalis		MG010092.1	EU532095.1	MG010110.1	FJ664504.1	EU532131.1	MG010060.1	MG010074.1	EU532100.1	U51087.1
		D. willistoni		BK006338.1	BK006338.1	BK006338.1	FJ664508.1	MG010088.1	MG010061.1	MG010075.1	EU532112.1	U51070.1
	bocainensis	*D. bocainensis*		OQ351288	OQ357798	OQ357799	OQ357802	N/A	OQ357803	OQ357805	OQ357804	OQ357806
		D. capricorni		OQ351289	OQ357800	OQ357801	N/A	AY335196.1	MG010068.1	MG010082.1	EU532113.1	U51092.1
		D. fumipennis		MG010101.1	EU532081.1	MG010119.1	FJ664509.1	EU532133.1	N/A	MG010083.1	EU532115.1	N/A
		D. nebulosa		EU493640.1	EU493769.1	EU494152.1	AF293717.1	U95275.1	MG010069.1	MG010084.1	EU532116.1	U51090.1
		D. sucinea		MG010103.1	EU532094.1	MG010121.1	FJ664510.1	AY335197.1	MG010070.1	MG010085.1	EU532114.1	U51091.1
saltans	*sturtevanti*	*D. sturtevanti*		AF050751.1	AF045081.1	N/A	N/A	AY335201.1	MG010071.1	MG010086.1	EU532117.1	AY335219.1
	saltans	*D. saltans*		MG010104.1	HQ110562.1	MG010122.1	N/A	AB026533.1	N/A	N/A	N/A	AY335218.1
melanogaster	*melanogaster*	*D. melanogaster*		KT174474.1	KT174474.1	KT174474.1	X04695.1	Z00030.1	X04661.1	NM_001111012.3	NM_169233.2	NM_080317.2
virilis	*virilis*	*D. virilis*		BK006340.1	BK006340.1	BK006340.1	AF293729.1	DQ471668.1	AF293749.1	EU514470.3	EF635085.1	X13877.1

N/A = sequence not available.

Table S2. Statistics of gene concordance (gCF) and site concordance (sCF) factors analysis for the ASTRAL species tree.

ID	gCF	gCF_N	gDF1	gDF1_N	gDF2	gDF2_N	gDFP	gDFP_N	gN	sCF	sCF_N	sDF1	sDF1_N	sDF2	sDF2_N	sN
16	29.57	207	41.86	293	28	196	0.57	4	700	NA	NA	NA	NA	NA	NA	NA
17	91.17	547	0.17	1	0	0	8.67	52	600	59.57	132.07	23.02	50.93	17.41	39.2	222.2
18	100	300	0	0	0	0	0	0	300	54.48	60.55	18.26	19.73	27.27	30.1	110.38
19	35.5	213	15.5	93	9.83	59	39.17	235	600	42.02	61.28	31.25	48.24	26.73	41.56	151.08
20	66.67	400	2	12	11.5	69	19.83	119	600	49.77	69.55	28.11	42.66	22.12	29.83	142.04
21	79.8	399	0	0	17.4	87	2.8	14	500	85.81	103.01	4.64	5.76	9.55	13.24	122.01
22	31.57	221	13	91	21.29	149	34.14	239	700	31.54	45.48	34.23	46.61	34.23	44.12	136.21
23	54	270	1	5	9	45	36	180	500	43.66	42.66	27.27	25.62	29.06	26.68	94.96
24	100	700	0	0	0	0	0	0	700	81.65	134.95	8.64	14.73	9.7	16.77	166.45
25	80	560	1.71	12	4	28	14.29	100	700	49.52	46.49	27.76	26.46	22.73	20.9	93.85
26	68.29	478	2.71	19	14.29	100	14.71	103	700	37.04	26.14	40.71	29.02	22.25	15.88	71.04
27	99.43	696	0.29	2	0.14	1	0.14	1	700	72.39	44.56	10.77	6.54	16.84	9.62	60.72

ID: Branch ID

gCF: Gene concordance factor (=gCF_N/gN %)

gCF_N: Number of trees concordant with the branch

gDF1: Gene discordance factor for NNI-1 branch (=gDF1_N/gN %)

gDF1_N: Number of trees concordant with NNI-1 branch

gDF2: Gene discordance factor for NNI-2 branch (=gDF2_N/gN %)

gDF2_N: Number of trees concordant with NNI-2 branch

gDFP: Gene discordance factor due to polyphyly (=gDFP_N/gN %)

gDFP_N: Number of trees decisive but discordant due to polyphyly

gN: Number of trees decisive for the branch

sCF: Site concordance factor averaged over 100 quartets (=sCF_N/sN %)

sCF_N: sCF in absolute number of sites

sDF1: Site discordance factor for alternative quartet 1 (=sDF1_N/sN %)

sDF1_N: sDF1 in absolute number of sites

sDF2: Site discordance factor for alternative quartet 2 (=sDF2_N/sN %)

sDF2_N: sDF2 in absolute number of sites

sN: Number of informative sites averaged over 100 quartets

LocalPP: ASTRAL-III local posterior probability

Length: Branch length

gEF_p: P-value for equal frequencies of gene trees between discordant trees

sEF_p: P-value for equal frequencies of sites between discordant trees

Table S3. Statistics of gene concordance (gCF) and site concordance (sCF) factors analysis for the simulated tree, with a hypothetical monophyletic *bocainensis* subgroup.

ID	gCF	gCF_N	gDF1	gDF1_N	gDF2	gDF2_N	gDFP	gDFP_N	gN	sCF	sCF_N	sDF1	sDF1_N	sDF2	sDF2_N	sN
16	28	196	29.57	207	41.86	293	0.57	4	700	NA	NA	NA	NA	NA	NA	NA
17	91.17	547	0	0	0	0	8.83	53	600	58.97	137.19	23.34	54.39	17.7	41.94	233.52
18	35.5	213	3.33	20	7.17	43	54	324	600	41.56	65.31	27.88	45.66	30.55	49.23	160.2
19	100	700	0	0	0	0	0	0	700	78.52	134.62	10.69	20.77	10.79	19.23	174.62
20	80	560	4	28	1.71	12	14.29	100	700	49.03	46.96	21.72	20.49	29.25	27.86	95.31
21	68.29	478	14.29	100	2.71	19	14.71	103	700	36.24	26.95	22.04	16.7	41.72	31.31	74.96
22	99.43	696	0.29	2	0.14	1	0.14	1	700	72.68	46.92	10.97	6.94	16.35	9.73	63.59
23	13	91	21.29	149	31.57	221	34.14	239	700	34.74	46.2	32.79	41.87	32.47	44.82	132.89
24	66.67	400	0	0	0	0	33.33	200	600	48.09	55.11	26.24	28.34	25.68	30.84	114.29
25	79.8	399	0	0	17.4	87	2.8	14	500	86.28	107.71	4.36	5.59	9.36	13.02	126.32
26	54	270	7.6	38	0	0	38.4	192	500	45.1	44.88	29.92	27.74	24.98	22.58	95.2
27	100	300	0	0	0	0	0	0	300	54.52	59.79	18.59	19.89	26.89	29.34	109.02

ID: Branch ID

gCF: Gene concordance factor (=gCF_N/gN %)

gCF_N: Number of trees concordant with the branch

gDF1: Gene discordance factor for NNI-1 branch (=gDF1_N/gN %)

gDF1_N: Number of trees concordant with NNI-1 branch

gDF2: Gene discordance factor for NNI-2 branch (=gDF2_N/gN %)

gDF2_N: Number of trees concordant with NNI-2 branch

gDFP: Gene discordance factor due to polyphyly (=gDFP_N/gN %)

gDFP_N: Number of trees decisive but discordant due to polyphyly

gN: Number of trees decisive for the branch

sCF: Site concordance factor averaged over 100 quartets (=sCF_N/sN %)

sCF_N: sCF in absolute number of sites

sDF1: Site discordance factor for alternative quartet 1 (=sDF1_N/sN %)

sDF1_N: sDF1 in absolute number of sites

sDF2: Site discordance factor for alternative quartet 2 (=sDF2_N/sN %)

sDF2_N: sDF2 in absolute number of sites

sN: Number of informative sites averaged over 100 quartets

LocalPP: ASTRAL-III local posterior probability

Length: Branch length

gEF_p: P-value for equal frequencies of gene trees between discordant trees

sEF_p: P-value for equal frequencies of sites between discordant trees

Supplementary Material2 – Results of DAMBE saturation tests

Alcohol dehydrogenase (*Adh*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites	0.2876
Mean H	0.3403
Standard Error	0.0247
Hmax	1.8395
Iss	0.1850
Iss.c	0.6981
T	20.8091
DF	288
Prob (Two-tailed)	**0.0000**
95% Lower Limit	0.1365
95% Upper Limit	0.2335

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.4999
T	12.7691
DF	288
Prob (Two-tailed)	**0.0000**
95% Lower Limit	0.1365
95% Upper Limit	0.2335

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Alpha-methyldopa hypersensitive (AMD)

est of substitution saturation (ia et al. 00 ia and emey 00)

esting hether the observed Iss is significantly lo er than Iss.c.

art I. or a symmetrical tree.

rop. invar. sites		0.
Mean		0. 1
Standard Error		0.01
max		1.
Iss	0.	1
Iss.c	0.	
	.11	
D		
rob (o-tailed)		0.0000
o er imit		0.1
pper imit		0.

art II. or an extreme asymmetrical (and generally very unli ely) tree.

Iss.c	0.	11
	0.01	
D		
rob (o-tailed)		0.0000
o er imit		0.1
pper imit		0.

Interpretation of results

	Significant Difference	
	es	o
Iss Iss.c	ittle	Substantial
	saturation	saturation
Iss Iss.c	seless	ery poor
	se uences	for phylogenetics

Cytochrome c oxidase subunit I (*Cox1*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.3290
Mean H		0.5460
Standard Error		0.0156
Hmax		1.7468
Iss	0.3126	
Iss.c	0.7594	
T	28.7114	
DF	549	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2820
95% Upper Limit		0.3431

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5548	
T	15.5635	
DF	549	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2820
95% Upper Limit		0.3431

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Cytochrome c oxidase subunit II (*Cox2*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.6387
Mean H		0.5272
Standard Error		0.0270
Hmax		1.6970
Iss	0.3107	
Iss.c	0.7338	
T	15.6696	
DF	227	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2575
95% Upper Limit		0.3638

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5171
T	7.6429
DF	227
Prob (Two-tailed)	**0.0000**
95% Lower Limit	0.2575
95% Upper Limit	0.3638

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Cytochrome B (*cytb*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.3245
Mean H		0.4661
Standard Error		0.0136
Hmax		1.6578
Iss	0.2811	
Iss.c	0.7426	
T	33.9972	
DF	481	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2545
95% Upper Limit		0.3078

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5457	
T	19.4892	
DF	481	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2545
95% Upper Limit		0.3078

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Dopa decarboxylase (*Ddc*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.5617
Mean H		0.7831
Standard Error		0.0149
Hmax		1.8420
Iss	0.4251	
Iss.c	0.7675	
T	22.9471	
DF	409	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.3958
95% Upper Limit		0.4544

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5672	
T	9.5230	
DF	409	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.3958
95% Upper Limit		0.4544

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

***Hunchback* (*Hb*)**

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.6230
Mean H		0.8110
Standard Error		0.0202
Hmax		1.8482
Iss	0.4388	
Iss.c	0.7353	
T	14.6589	
DF	241	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.3990
95% Upper Limit		0.4786

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5188	
T	3.9525	
DF	241	
Prob (Two-tailed)		**0.0001**
95% Lower Limit		0.3990
95% Upper Limit		0.4786

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

Male fertility factor *kl-3* (*kl-3*)

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.5579
Mean H		0.6632
Standard Error		0.0311
Hmax		1.7757
Iss	0.3735	
Iss.c	0.7068	
T	10.7085	
DF	194	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.3121
95% Upper Limit		0.4349

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5182
T	4.6492
DF	194
Prob (Two-tailed)	**0.0000**
95% Lower Limit	0.3121
95% Upper Limit	0.4349

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

***period* (*per*)**

Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

Testing whether the observed Iss is significantly
lower than Iss.c.

Part I. For a symmetrical tree.

Prop. invar. sites		0.3087
Mean H		0.5531
Standard Error		0.0150
Hmax		1.8378
Iss	0.3009	
Iss.c	0.7774	
T	31.8644	
DF	770	
Prob (Two-tailed)		**0.0000**
95% Lower Limit		0.2716
95% Upper Limit		0.3303

Part II. For an extreme asymmetrical (and generally very
unlikely) tree.

Iss.c	0.5833
T	18.8864
DF	770
Prob (Two-tailed)	**0.0000**
95% Lower Limit	0.2716
95% Upper Limit	0.3303

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	Little saturation	Substantial saturation
Iss > Iss.c	Useless sequences	Very poor for phylogenetics

CAPÍTULO III

Patterns of genome size evolution *versus* fraction of repetitive elements in *statu nascendi* species: the case of the *willistoni* subgroup of *Drosophila* (Diptera, Drosophilidae)

Henrique R.M. Antoniolli

Maríndia Deprá

Vera L.S. Valente

Manuscrito publicado no periódico *Genome* (ISSN: 1480-3321)

doi: 10.1139/gen-2022-0073

Genome
An International Journal
for Genetics and Genomics

Article

Patterns of genome size evolution versus fraction of repetitive elements in *statu nascendi* species: the case of the *willistoni* subgroup of *Drosophila* (Diptera, Drosophilidae)

Henrique R.M. Antoniolli, Maríndia Deprá, and Vera L.S. Valente

Laboratório de Drosophila, Programa de Pós-graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

Corresponding author: **Maríndia Deprá** (email: marindiadepra@gmail.com)

Abstract

Genome size evolution is known to be related with transposable elements, yet such relation in incipient species remains poorly understood. For decades, the *willistoni* subgroup of *Drosophila* has been a model for evolutionary studies because of the different evolutionary stages and degrees of reproductive isolation its species present. Our main question here was how speciation influences genome size evolution and the fraction of repetitive elements, with a focus on transposable elements. We quantitatively compared the mobilome of four species and two subspecies belonging to this subgroup with their genome size, and performed comparative phylogenetic analyses. Our results showed that genome size and the fraction of repetitive elements evolved according to the evolutionary history of these species, but the content of transposable elements showed some discrepancies. Signals of recent transposition events were detected for different superfamilies. Their low genomic GC content suggests that in these species transposable element mobilization might be facilitated by relaxed natural selection. Additionally, a possible role of the superfamily DNA/*TcMar-Tigger* in the expansion of these genomes was also detected. We hypothesize that the undergoing process of speciation could be promoting the observed increase in the fraction of repetitive elements and, consequently, genome size.

Key words: insects, mobilome, phylogenetic signal, speciation, TEs landscape

Introduction

The evolution of genome size (GS), or C-value, across the tree of life has been a puzzling question in genome biology for decades. For instance, the C-value of the protozoan genus *Amoeba* is about 200 times larger than that of humans, while within Eukarya it ranges 200 000-fold (Gregory 2004). Many hypotheses attempt to understand this variation, addressing the correlation of GS with, for instance, cell and body sizes (Beaulieu et al. 2008), metabolism rates (Waltari and Edwards 2002), developmental complexity (Gregory 2002), and evolutionary history of species (Jeffery et al. 2017). In other words, these hypotheses address factors that act by relaxing the maximum limits of GS, which enables its expansion. The main mechanism causing genome expansions, however, appears to be the activity of transposable elements (TEs) (Kidwell 2002; Gregory and Johnston 2008; Canapa et al. 2015; Sessegolo et al. 2016; Cong et al. 2022; Heckenhauer et al. 2022), which comprises factors that actively add new DNA sequences.

TEs are linear DNA sequences that belong to the repetitive fraction of genomes and are able to move within and even "jumping" to another species' DNA content (new reviews in Cordeiro et al. 2019; Melo and Wallau 2020; Wells and Feschotte 2020; Zhang et al. 2020; Palazzo et al. 2021; Ahmad et al. 2022). A portion of virtually all genomes is made up of TEs—for example, up to 80% in maize (Schnable et al. 2009), 44% in humans (Mills et al. 2007), and 50% in insects, such as orthopterans (Camacho et al. 2015) and dipterans (Nene et al. 2007). These sequences are divided into two major groups: (i) retrotransposons, or Class I; and (ii) DNA transposons, or Class II (Finnegan 1989; Wicker et al. 2007). The first class is known to transpose by synthesizing an intermediary mRNA molecule, while the latter transposes either by cutting and pasting itself elsewhere in the genome, or by copying itself directly into DNA (Arkhipova 2017).

The mechanism of replicative transposition amplifies the copy number (CN) of a given TE, thus increasing the total amount of DNA within a cell (Su et al. 2018). Additionally, TEs are known to cause chromosomal rearrangements, including deletions, duplications, translocations, and inversions by ectopic recombination (Delprat et al. 2009; González and Petrov 2012), changing the chromosome architecture (Ren et al. 2018) and thus changing the C-value (Kidwell 2002; Bourque

et al. 2018). The activity or presence of TEs and their implications on and relationships with GS evolution are well described in the literature, whereas the effect of that relationship in actively speciating taxa remains poorly understood (Kraaijeveld 2010).

For more than half of a century, the dipterans belonging to the *willistoni* group of the genus *Drosophila* offer a unique window to observe many biological phenomena under the process of speciation—for example, from chromosomal rearrangements (review in Rohde and Valente 2012) to courtship behavior (Ritchie and Gleason 1995), genome methylation (Garcia et al. 2007), reproductive isolation (Dobzhansky and Mayr 1944), and shifts in levels of GC-content (Tarrío et al. 2000). In fact, the *willistoni* group and its sister *saltans* group possess species with some of the known lowest GC-contents in *Drosophila* (Rodríguez-Trelles et al. 2000; Tarrío et al. 2001; Vicario et al. 2007). Both groups comprise the neotropical clade of the subgenus *Sophophora* (O'Grady and Kidwell 2002), and the *willistoni* group is further divided into the *alagitans* (6 species), *bocainensis* (12 species), and *willistoni* (6 species) subgroups (Bächli 2022). Additionally, both *saltans* and *willistoni* groups comprise a clade sister to the *Lordiphosa* genus, which is phylogenetically placed within the *Sophophora* subgenus (Kim et al. 2021; Suvorov et al. 2022). In fact, the evolutionary history of *Drosophila* reveals that its species actually form a paraphyletic lineage, with several genus branching off within (Yassin 2013; O'Grady and DeSalle 2018).

The main reason for such an outstanding place in evolutionary biology are the *in statu nascendi* species of the *willistoni* subgroup (Dobzhansky and Spassky 1959). In other words, it comprises sibling species in different taxonomic levels and degrees of reproductive isolation (Robe et al. 2010; Zanini et al. 2018). For instance, fertile hybrids have been experimentally produced (review in Winge and Cordeiro 1963), while other crosses do not produce hybrids, or the offspring die at larval stage (Winge 1965). The most remarkable example is *D. paulistorum*, a complex of six subspecies able to interbreed and produce hybrid offspring (Dobzhansky and Spassky 1959). In fact, this complex started to diverge around 2 million years ago, and some of these subspecies diverged around 300 000 years ago—the case of *D. paulistorum andeanbrazilian* and *D. paulistorum orinocan* (see a chronogram in Zanini et al. 2018).

According to Gregory and Johnston (2008), non-coding DNA is the main driving force linked to GS variability in *Drosophila*. However, distinct patterns in this relationship were observed between the *Drosophila* and *Sophophora* subgenera: while satellite DNA is highly associated with GS in the former, TEs are the main correlate with GS in the latter (De Lima and Ruiz-Ruano 2022). This study, nonetheless, did not include the *willistoni* group. Given this, here we studied how the speciation process may influence GS evolution and fraction of repetitive elements (FRE), focusing on TEs, in insects. We used the (sub)species of the *willistoni* group as a case study by comparing these traits under a quantitative approach. We provide a general overview on the evolution of these traits, and our main questions included: how GS and FRE evolved in this subgroup; how strongly are both related with the evolutionary history of those species; and how the TEs composition differs between them.

Material and methods

Genomic data acquisition

We retrieved from GenBank genome assemblies of seven species or subspecies belonging to the *bocainensis* and *willistoni* subgroups of *Drosophila*—*D. equinoxialis*, *D. insularis*, *D. paulistorum* ssp., *D. paulistorum andeanbrazilian*, *D. sucinea*, *D. tropicalis*, and *D. willistoni* (see details on taxonomy and accession nos. on Table 1). Genome assemblies of species belonging to the *saltans* group and the *Lordiphosa* genus were also downloaded and included as an outgroup in the downstream analyses. Among the available assemblies from different strains of *D. willistoni*, we selected the first assembled genome for the species, from the isolineage Gd-H4-1. This lineage was inbred under laboratory conditions for nine generations aiming to reduce the typical chromosomal rearrangements found in this species (Rohde and Valente 2012) and thus to be better suitable for NGS (Drosophila 12 Genomes Consortium et al. 2007). All genomes included in this study were assembled with long reads, and their completeness was assessed through the search of 3285 BUSCO v.5.0.0 markers (Manni et al. 2021) from the "Diptera_odb10" database.

Statistical analyses on interspersed repeats landscapes and TEs

We ran RepeatMasker 4.1.2 (Smit et al. 2022) using rmblast in the genome assemblies of the *saltans* and *willistoni* subgroups, with the combined Dfam and RepBase-20181026 databases. Then, we produced interspersed repeats landscape graphs and estimated the FRE with the "calcDivergenceFromAlign.pl" and "createRepeatLandscape.pl" scripts included in the RepeatMasker package.

Next, we parsed the RepeatMasker output of the genome assemblies of the *willistoni* group with the script "One Code to Find them All" (Bailly-Bechet et al. 2014) to obtain a table with the CN and total length (TL) of each TE superfamily found in each genome. Our aim with this analysis is to quantitatively evaluate how similar the content of TEs is between the species of the *willistoni* subgroup. In this case, *D. sucinea* was included to serve as the outgroup, as it belongs to the *bocainensis* subgroup. We set the flag—*strict* to select hits based on the "80–80–80 rule" of Wicker et al. (2007). This table was split into two datasets: (i) a dataset containing the CN of each TE superfamily—hereafter called CN; and (ii) a dataset containing the TL, in base pairs (bp), of these superfamilies—hereafter called TL (Tables S1 and S2, respectively). We chose to analyze these two variables because both of them might be linked with increasing or decreasing the overall size of genomes. Both datasets were log (ln) transformed and submitted to principal component analysis, with the *PCAtools* package (Blighe and Lun 2022) in R 4.1.2 (R Core Team 2021). Superfamilies were classified according to Kapitonov and Jurka (2008).

Table 1. Information on taxonomy, GenBank accession numbers, genome size (in megabases), and fraction of repetitive elements of each species included in this study.

Genus	Species group	Species	Subspecies	Accession No.	Genome size (Mb)	GC%	Repetitive elements	BUSCO**
Drosophila	*obscura*	*Drosophila pseudoobscura*		GCA_018904455.1	N/A	45.02%	N/A	98.6%
	saltans	*D. neocordata*		GCA_018903615.1	178.96*	35.02%	23.33%	99.3%
		D. prosaltans		GCA_018151275.1	211.61*	36.07%	26.34%	99.1%
		D. saltans		GCA_018903575.1	204.12*	36.17%	30.51%	99.2%
		D. sturtevanti		GCA_018150375.1	174.2*	34.44%	21.44%	98.8%
	willistoni	*D. sucinea*		GCA_018150745.1	205.38	37.70%	17.32%	98.3%
		D. equinoxialis		GCA_018150345.1	273.84	37.66%	33.02%	97.3%
		D. insularis		GCA_018903935.1	202.47*	36.89%	30.03%	99.2%
		D. paulistorum	*andeanbrazilian*	GCA_018151315.1	221.3*	37.87%	37.89%	99.0%
		D. paulistorum	ssp.	GCA_018152135.1	231.02*	38.19%	41.02%	99.1%
		D. tropicalis		GCA_018151085.1	225.1*	37.97%	32.88%	98.2%
		D. willistoni		GCA_018902025.2	220.05	37.25%	29.13%	98.9%
Lordiphosa		*Lordiphosa clarofinis*		GCA_018904275.1	540.87*	38.19%	19.33%	98.0%
		L. collinella		GCA_018904265.1	389.29*	39.01%	22.78%	96.4%
		L. magnipectinata		GCA_018904285.1	542.11*	38.23%	21.99%	97.0%
		L. mommai		GCA_018904225.1	322.38*	36.17%	22.32%	97.0%
		L. stackelbergi		GCA_018904235.1	468.9*	38.05%	21.54%	76.9%

*Mean between estimates based on single copy genes coverage and k-mer length performed by Kim et al. (2021).
**Percentage of complete single copy genes found in the assemblies with BUSCO, based on the Diptera database with 3285 genes. N/A = not available; *D. pseudoobscura* served as an outgroup for the phylogenetic tree inference.

Statistical analyses on GC-content

For the species of the *saltans* and *willistoni* groups, we recovered the percentage of GC-content from the outputs of the RepeatMasker analysis. To confirm that both groups present low levels of GC-content, we collected estimates for several species of *Drosophila* performed by Bronski et al. (2020) and gathered by Seetharam and Stuart (2013). We also included the estimation for *D. pseudoobscura* (included in the outgroup in the section below). We then performed a Wilcoxon–Mann–Whitney test, given that the data was not normally distributed—according to Shapiro–Wilk tests. Data and scripts are deposited in GitHub (https://github.com/henriqueantoniolli/patterns-of-genome-size-on-willistoni).

Statistical analyses on ancestral state reconstructions

The single copy genes found by BUSCO for assessing the completeness of *Drosophila* and *Lordiphosa* assemblies also served as input to a pipeline written by McGowan (2020). Briefly, each orthologous gene marked as complete and single copy had its nucleotide sequences aligned with MUSCLE v3.8.15 (Edgar 2004), trimmed with TrimAl (Capella-Gutiérrez et al. 2009), and translated into amino acids. All genes found in common between the genomes were concatenated into a single matrix. The best substitution model was chosen automatically by IQ-TREE 2 (Minh et al. 2020) based on AIC scores (setting *-m TEST—merit AIC*), and 1000 ultrafast bootstrap replicates (UFboot) were employed to estimate branch support. Branch lengths were adjusted by estimating divergence times with treePL (Smith and O'Meara 2012), using as calibration point the 95% confidence interval for the most recent common ancestor between the *saltans* and *willistoni* groups (13.48–19.78 million years), estimated by Suvorov et al. (2022).

The time-calibrated supermatrix tree was used for the downstream analyses. We mapped the evolutionary history of both the GS and the FRE of the genomes of the *willistoni* (ingroup) and *saltans* (outgroup) groups. In the first ancestral state reconstruction, we used GS estimations based on flow cytometry available at the Animal GS Database (Gregory 2022) and converted the data from picograms (pg) to megabases (Mb), according to the formula 1 pg = 978 Mb (Doležel et al. 2003). For those species with more than one estimation, we calculated their mean. For those species with no data available, we used the estimations performed by Kim et al. (2021) (Table 1). In this case, the GS is a mean between estimations based on single copy genes coverage and based on k-mers length. This strategy was employed to minimize possible biases in both methods—as for some species the estimates varied in a large range (e.g., *D. insularis*; 233.6 and 171.31Mb, respectively). In the second ancestral state reconstruction, we used the FRE estimated by the RepeatMasker analysis, as described in the subsections above. We then used the function *fastAnc* from the *phytools* package (Revell 2012) in R 4.1.2 to map both GS and FRE onto the supermatrix tree.

Both data on GS and FRE were log (ln) transformed and submitted to a phylogenetic generalized least squares (PGLS) test with the *pgls* function from the *caper* package (Orme et al. 2018) in R 4.1.2. In this analysis, we tested the direction of the relationship between GS and FRE with their underlying phylogenetic relationships. The *delta*, *lambda*, and *kappa* values of branch length transformation were optimized by maximum likelihood, according to the data and model. Both models of direction (GS∼FRE and FRE∼GS) were tested. Tests of phyloge-

netic signals were performed with the whole set of *Drosophila* and *Lordiphosa* species (a total of 16 species) using the *phylosig* function from the *phytools* package, which estimates Pagel's λ and Bloomsberg's K values.

Results

Statistical analyses

The largest GS among the *saltans* and *willistoni* groups belongs to *D. equinoxialis*, while the smallest belongs to *D. sucinea* and *D. sturtevanti* (Table 1); approximately 273.84, 202.47, and 174.2, respectively (Kim et al. 2021). The Wilcoxon–Mann–Whitney test showed that the *saltans* and *willistoni* groups have statistically significant lower GC-content ($W = 470$; p value = 5.3e-07) than the other species of *Drosophila*.

The analyses conducted in RepeatMasker for reconstructing the interspersed repeats landscape resulted in distinct patterns of distribution of Kimura substitution levels. The interspersed repeats landscape of *D. sturtevanti* and *D. sucinea* (Fig. 1) and *D. neocordata*, *D. prosaltans*, and *D. saltans* (Fig. S1), showed a bimodal distribution (Fonseca et al. 2019). Among the species of the *willistoni* subgroup, the landscapes were similar and presented less sharp bimodal distributions that resembled the "L" shape (Fig. 1). Low levels of Kimura substitution represent less variation between TEs copies, then it is expected that these copies had recent events of transposition—representing a "L" shape on the landscape. On the other hand, a bimodal distribution implies that most TEs copies found in a genome are no longer active. The fraction of the genome that corresponds to repetitive sequences for the species of the *saltans* group and *D. sucinea* was generally low—except for *D. saltans*, whose fraction was estimated as 30.51% (Table 1). The results obtained for the *willistoni* subgroup, however, were quite higher: from 30% in *D. insularis* and 29.1% in *D. willistoni*, to 41% in *D. paulistorum* ssp. and 38.9% in *D. paulistorum andeanbrazilian* (Fig. 1, Table 1). Simple repeats (such as microsatellites) maintained similar and lower portions (~4.5%) than TEs in the genome in the *willistoni* subgroup (Fig. 1). In the outgroups and *D. sucinea*, simple repeats are the larger portion in the FRE (Fig. 1; Fig. S1). On the other hand, the portion of TEs superfamilies vary and seems to be higher than in the outgroup species, including the *willistoni* group species *D. sucinea*.

The principal component analysis showed that *D. sucinea* separated from the *willistoni* subgroup, mainly in the first principal component, on both CN (Fig. 2A) and TL (Fig. 2B) datasets—CN is the dataset containing the CN of each TE superfamily and TL is a dataset containing the TL, in base pairs (bp), of these superfamilies. The superfamilies DNA/*Sola-1*, DNA/*TcMar-Tigger*, LINE/*CR1*, LINE/*Penelope*, LINE/*R1*, and LTR/*Pao* are the most contributors for the variation found in CN dataset (Fig. S3). Nonetheless, the most contributors in the TL datasets were DNA/*hAT-Charlie*, DNA/*Merlin*, DNA/*MULE-NOF*, and DNA/*TcMar-Tigger* (Fig. S4). The first and second principal component explains, for CN and TL datasets, 60.77% and 16.64% (total of 77.41%), and 47.56% and 23.49% (total of 71.05%) of the variation, respectively.

Ancestral state reconstructions and phylogenetic signals

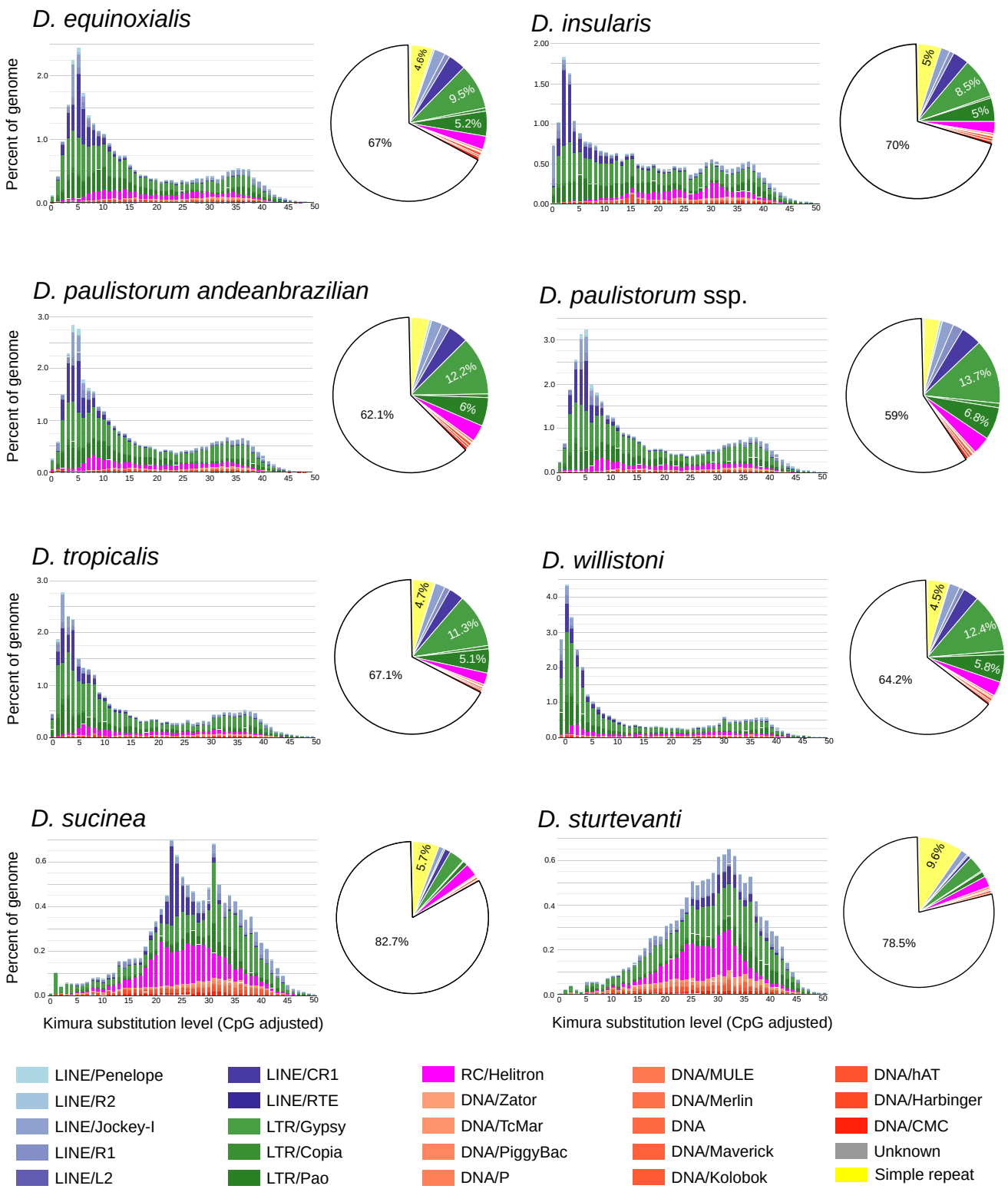
A total of 2928 (from 3285) orthologous genes were found by BUSCO as complete and single copy in all the species studied here, and were added into the supermatrix. The best substitution model estimated by IQ-TREE 2 was JTT + F + I + G4. Branches presented high ultrafast bootstrap support (UFboot = 100), and the topology (Fig. S2) agreed with previous phylogenetic reconstructions (Gleason et al. 1998; Tarrío et al. 2000; Robe et al. 2010), including the divergence times (Zanini et al. 2018). The ancestral state reconstructions of GS (Fig. 3A) and FRE (Fig. 3B) uncovered similar patterns, as species closer to the root showed lower values for both traits, while derived species presented higher values. Pagel's λ was close to one and statistically significant for both GS and FRE ($\lambda = 0.95973$, p value = 8.84844e-07, and $\lambda = 0.937491$, p value = 0.00195, respectively). On the other hand, Bloomsberg's K was, respectively, higher and lower than 1 for GS ($K = 1.73699$, p value = 0.001) and FRE ($K = 0.81763$, p value = 0.002). Statistically significant (p value = 0.003076) and positive relationship ($b = 0.204039$) between GS and FRE was only found when the latter is the explanatory variable (GS ~ FRE) in the PGLS analysis (Table S3), i.e., GS increases as FRE increases. The model FRE~GS (Table S4) was not statistically significant (p value = 0.8821).

Discussion

Although some of our results might speculate patterns of genome evolution and FRE in the sister *saltans* group of *Drosophila*, here we focused to understand them in the *willistoni* subgroup. The latter comprises *in statu nascendi* species, which means the process of speciation is not entirely complete and those entities may still interbreed and generate hybrid offspring (Dobzhansky and Spassky 1959). Divergence times recently estimated by Zanini et al. (2018) further support that the subspecies of the *D. paulistorum* complex have not yet completed the speciation process, as most of them diverged less than one million years ago. It is expected that closely related species—in the case of the *willistoni* subgroup, we may call these as sibling species given their short divergence times—would present similar GS and content of repetitive or TEs. Interestingly, our results showed quite the opposite, especially considering the latter.

Signals of recent transpositions were detected in the interspersed repeats landscapes of species belonging to the *willistoni* subgroup (Fig. 1), as peaks in Kimura distance closer to 0 (i.e., the "L" shape) indicate that the copies found within a genome are similar between each other (Fonseca et al. 2019). Bursts of transposition might be associated with speciation events, since hybridization and stressful environmental conditions faced by species while expanding into new areas may result in the breakdown of epigenetic control of TEs, triggering their mobilization and amplification (Gregory 2001; Rebollo et al. 2010). Even though most of the crossings produce sterile offspring, interspecific hybridizations have been reported in the *willistoni* subgroup (review in Winge and Cordeiro 1963)—as the case of *D. equinoxialis* and *D. tropicalis*;

Fig. 1. Interspersed repeat landscape and genome fraction estimates of each genome included in this study.



D. paulistorum and *D. insularis*; *D. paulistorum* and *D. equinoxialis*; and *D. willistoni* and *D. insularis* (Winge 1965). Various levels of hybridization between subspecies were also reported for the *D. paulistorum* complex (Dobzhansky et al. 1969). Interestingly, approximately 2.75% of the *D. willistoni* and 0.7% of the *D. insularis* genomes showed zero levels of Kimura sub-

Fig. 2. Principal component analysis biplot on (A) copy number and (B) total length of each superfamily dataset. Axes X and Y represent, respectively, principal components 1 and 2.

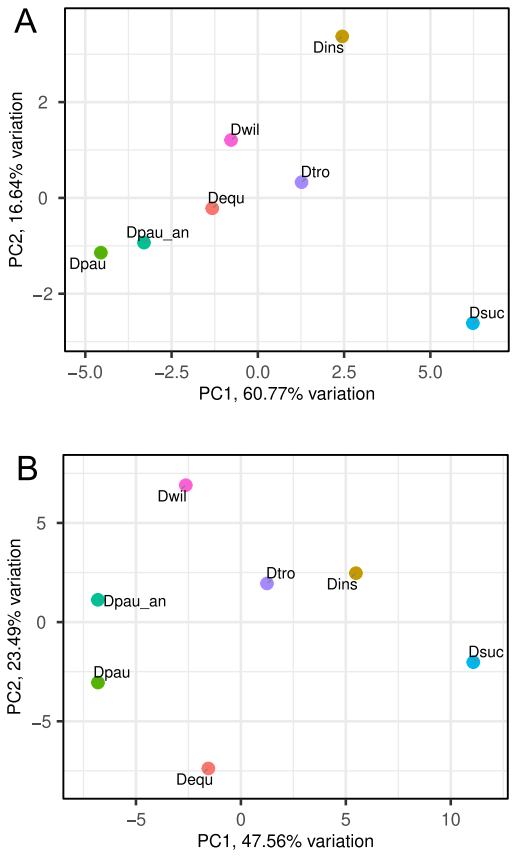
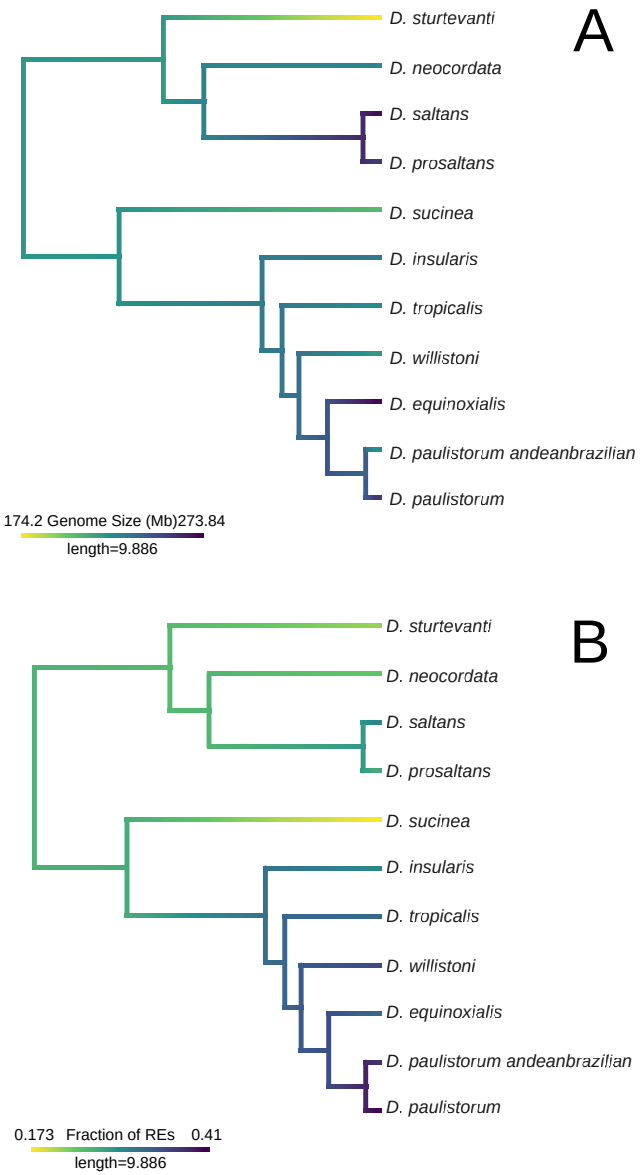


Fig. 3. Ancestral state reconstructions of (A) genome size (GS) and (B) fraction of repetitive elements (REs).



stitution. Regarding the first species, the majority of them are LTR retrotransposons; on the latter, however, more than a half are LINE retrotransposons—an indication that TEs were distinctly mobilized in each genome. The distinct fractions within the FRE, regarding simple repeats and different superfamilies of TEs, reinforce the hypothesis that the latter were recently transposed in the *willistoni* subgroup as their portions are larger in these species than in the outgroups.

Although the species of the *willistoni* subgroup were spatially grouped in both principal component analysis biplots, a few superfamilies of TEs also showed to be distinctly present in these genomes. The majority of TEs superfamilies that are most different in CN and TL between genomes in the *willistoni* subgroup are, respectively, retrotransposons and DNA transposons (Fig. 2). Interestingly, DNA/*TcMar-Tigger* is the only superfamily to be present in both CN and TL biplots, suggesting a possible role of these elements in the expansion of these genomes, which should be further investigated.

The results of ancestral state reconstructions showed a general pattern of increasing values as species diverged. As the Pagel's λ values for both GS and FRE were close to 1, both traits evolved according to the evolutionary history of those species (Münkemüller et al. 2012; Molina-Venegas and Rodríguez 2017). However, contrasting scenarios are depicted from Blomsberg's K for GS and FRE. In this phylogeny, GS is estimated to be more similar between closely related species than a distant one—that is, differences are between clades (Blomsberg's $K > 1$), which agrees with the hypothesis of similarity between closely related species. Nonetheless, the FRE differ within clades ($K < 1$), which shows that although the content of repetitive elements followed the evolutionary history of these incipient species, it presents differences between them. This result is supported by the patterns observed in the interspersed repeat landscapes—in which different superfamilies of TEs were recently active in the genomes—and the differences observed in the principal component analysis biplots. The positive relationship between GS and FRE

detected by PGLS results supports that both traits are associated, which is already well established in the literature—for example, in the study of Kidwell (2002) with 12 eukaryote species; Gregory and Johnston (2008) and Sessegolo et al. (2016) in the genus *Drosophila*; and Heckenhauer et al. (2022) in trichopterans.

Additional evidence for the mobilization of TEs and the increase in GS of these species lies in their levels of GC-content. Natural selection is one of the main evolutionary mechanisms controlling the distribution and mobilization of TEs throughout their life cycle (Schaack et al. 2010; Bourque et al. 2018), and maintaining high levels of GC-content in *Drosophila* (Lawrie et al. 2013). The low levels of GC-content described for the *willistoni* and *saltans* groups (Tarrío et al. 2000, 2001; Vicario et al. 2007), and confirmed by our analysis, suggest a relaxation of natural selection in this clade. This phenomenon would facilitate the amplification of GS through the proliferation of TEs, as their presence (revealed by deleterious effects, such as inactivating or enhancing a gene expression or induce chromosomal rearrangements) would be less efficiently selected against.

We may conclude that the evolution of GS and the FRE reflect the evolutionary history of the *willistoni* subgroup. As a consequence, we may also speculate that the undergoing process of speciation led to larger FRE; indeed, until now, the highest FRE within *Drosophila* was found in *D. paulistorum* ssp. and *D. paulistorum andeanbrazilian* (Kim et al. 2021). Once speciation is complete, the activity of transposons reaches an equilibrium and results in a bell curve on interspersed repeats landscape, as described by Fonseca et al. (2019), and detected for *D. sturtevanti* and *D. sucinea* (Fig. 1). Similar results were found for cryptic species of rotifers (Stelzer et al. 2011), showing that high levels of GS amplification—even possible whole-genome duplications—may occur with speciation.

Acknowledgements

We thank MSc Ana Paula Tavares Costa for her valuable help with R scripts; Dr Adriana Ludwig, Dr Élgion Loreto and Dr Sebastián Pita for comments during the development of this study; and MSc Tuane Letícia Carvalho for helping with installation of software and comments on early versions of this manuscript. We also thank the two anonymous reviewers whose suggestions greatly improved our manuscript.

Article information

History dates

Received: 22 August 2022
Accepted: 22 April 2023
Accepted manuscript online: 25 April 2023
Version of record online: 18 May 2023

Copyright

Availability of data and materials

The data produced in this study are provided in the Supplementary Material and are deposited in GitHub: https://github.com/henriqueantoniolli/patterns-of-genome-size-on-willistoni

Author information

Author ORCIDs

Henrique R.M. Antoniolli https://orcid.org/0000-0002-6747-3641
Maríndia Deprá https://orcid.org/0000-0003-4568-1869

Author contributions

Conceptualization: HRMA
Data curation: HRMA
Formal analysis: HRMA, MD, VLSV
Funding acquisition: VLSV
Investigation: HRMA, MD, VLSV
Methodology: HRMA, MD
Software: HRMA
Supervision: MD, VLSV
Visualization: MD, VLSV
Writing – original draft: HRMA
Writing – review & editing: HRMA, MD, VLSV

Competing interests

The authors declare there are no competing interests.

Funding

This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under the scholarship No. 141319/2020-8 and research productivity grant No. 312781/2018-0. This study was also supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil.

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/gen-2022-0073.

References

Ahmad, A., Su, X., Harris, A.J., and Ren, Z. 2022. Closing the gap: horizontal transfer of mariner transposons between Rhus gall aphids and other insects. Biology **11**(5): 731. doi:10.3390/biology11050731. PMID: 35625459.

Arkhipova, I.R. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mobile DNA **8**(1): 19. doi:10.1186/s13100-017-0103-2. PMID: 29225705.

Bächli, G. 2022. TaxoDros. Available from https://www.taxodros.uzh.ch/ [accessed 11 August 2022].

Bailly-Bechet, M., Haudry, A., and Lerat, E. 2014. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mobile DNA **5**(1): 13. doi:10.1186/1759-8753-5-13.

Beaulieu, J.M., Leitch, I.J., Patel, S., Pendharkar, A., and Knight, C.A. 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. New Phytol. **179**(4): 975–986. doi:10.1111/j.1469-8137.2008.02528.x. PMID: 18564303.

Blighe, K., and Lun, A. 2022, August 3. PCAtools: Everything Principal Component Analysis. R. Available from https://github.com/kevinblighe/PCAtools [accessed 11 August 2022].

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. 2018. Ten things you should know about transposable elements. Genome Biol. **19**(1): 199. doi:10.1186/s13059-018-1577-z. PMID: 30454069.

Bronski, M.J., Martinez, C.C., Weld, H.A., and Eisen, M.B. 2020. Whole genome sequences of 23 species from the *Drosophila montium* species group (Diptera: Drosophilidae): a resource for testing evolutionary hypotheses. G3-Genes Genom. Genet. **10**(5): 1443–1455. doi:10.1534/g3.119.400959.

Camacho, J.P.M., Ruiz-Ruano, F.J., Martín-Blázquez, R., López-León, M.D., Cabrero, J., Lorite, P., et al. 2015. A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. Chromosoma **124**(2): 263–275. doi:10.1007/s00412-014-0499-0. PMID: 25472934.

Canapa, A., Barucca, M., Biscotti, M.A., Forconi, M., and Olmo, E. 2015. Transposons, genome size, and evolutionary insights in animals. CGR **147**(4): 217–239. Karger Publishers. doi:10.1159/000444429.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**(15): 1972–1973. doi:10.1093/bioinformatics/btp348. PMID: 19505945.

Cordeiro, J., Carvalho, T.L., Valente, V.L., da, S., and Robe, L.J. 2019. Evolutionary history and classification of Micropia retroelements in Drosophilidae species. PLoS ONE **14**(10): e0220539. Public Library of Science. doi:10.1371/journal.pone.0220539. PMID: 31622354.

Cong, Y., Ye, X., Mei, Y., He, K., and Li, F. 2022. Transposons and non-coding regions drive the intrafamily differences of genome size in insects. iScience **25**: 104873. doi:10.1016/j.isci.2022.104873. PMID: 36039293.

De Lima, L.G., and Ruiz-Ruano, F.J. 2022. In-depth satellitome analyses of 37 *Drosophila* species illuminate repetitive DNA evolution in the *Drosophila* genus. Genome Biol. Evol. **14**(5): evac064. doi:10.1093/gbe/evac064. PMID: 35511582.

Delprat, A., Negre, B., Puig, M., and Ruiz, A. 2009. The transposon Galileo generates natural chromosomal inversions in *Drosophila* by ectopic recombination. PLoS ONE **4**(11): e7883. Public Library of Science. doi:10.1371/journal.pone.0007883. PMID: 19936241.

Dobzhansky, T., and Mayr, E. 1944. Experiments on sexual isolation in *Drosophila*: I. Geographic strains of *Drosophila willistoni*. Proc. Natl. Acad. Sci. U.S.A. **30**: 238–244. doi:10.1073/pnas.30.9.238. PMID: 16588650.

Dobzhansky, T., and Spassky, B. 1959. *Drosophila paulistorum*, a cluster of species *in statu nascendi*. Proc. Natl. Acad. Sci. U. S. A. **45**(3): 419–428. doi:10.1073/pnas.45.3.419. PMID: 16590403.

Dobzhansky, T., Pavlovsky, O., and Ehrman, L. 1969. Transitional populations of *Drosophila paulistorum*. Evolution **23**(3): 482–492. [Society for the Study of Evolution, Wiley]. doi:10.2307/2406702. PMID: 28562925.

Doležel, J., Bartoš, J., Voglmayr, H., and Greilhuber, J. 2003. Letter to the editor. Cytometry A **51A**(2): 127–128. doi:10.1002/cyto.a.10013.

Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B.Drosophila 12 Genomes Consortium,, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450**(7167): 203–218. doi:10.1038/nature06341. PMID: 17994087.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340. PMID: 15034147.

Finnegan, D.J. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. **5**(4): 103–107. doi:10.1016/0168-9525(89)90039-5. PMID: 2543105.

Fonseca, P.M., Moura, R.D., Wallau, G.L., and Loreto, E.L.S. 2019. The mobilome of *Drosophila incompta*, a flower-breeding species: comparison of transposable element landscapes among generalist and specialist flies. Chromosome Res. **27**(3): 203–219. doi:10.1007/s10577-019-09609-x. PMID: 31119502.

Garcia, R.N., D'Ávila, M.F., Robe, L.J., Loreto, E.L.S., Panzera, Y., Heredia, F.O., and Valente, V.L.S. 2007. First evidence of methylation in the genome of *Drosophila willistoni*. Genetica **131**: 91–105. doi:10.1007/s10709-006-9116-3. PMID: 17205375.

Gleason, J.M., Griffith, E.C., and Powell, J.R. 1998. A molecular phylogeny of the *Drosophila willistoni* group: conflicts between species concepts? Evolution **52**(4): 1093–1103. doi:10.2307/2411239. PMID: 28565231.

González, J., and Petrov, D.A. 2012. Evolution of genome content: population dynamics of transposable elements in flies and humans. *In* Evolutionary Genomics: Statistical and Computational Methods, **Volume 1**. *Edited by* M. Anisimova. Humana Press, Totowa, NJ. pp. 361–383. doi:10.1007/978-1-61779-582-4_13.

Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc. **76**(1): 65–101. doi:10.1017/s1464793100005595. PMID: 11325054.

Gregory, T.R. 2002. Genome size and developmental complexity. Genetica **115**(1): 131–146. doi:10.1023/a:1016032400147. PMID: 12188045.

Gregory, T.R. 2004. Macroevolution, hierarchy theory, and the C-value enigma. pbio **30**(2): 179–202. The Paleontological Society. doi:10.1666/0094-8373(2004)030⟨0179:MHTATC⟩2.0.CO;2.

Gregory, T. 2022. Animal Genome Size Database. Available from http://www.genomesize.com/ [accessed 11 August 2022].

Gregory, T.R., and Johnston, J.S. 2008. Genome size diversity in the family Drosophilidae. Heredity **101**(3): 228–238. doi:10.1038/hdy.2008.49. PMID: 18523443.

Heckenhauer, J., Frandsen, P.B., Sproul, J.S., Li, Z., Paule, J., Larracuente, A.M., et al. 2022. Genome size evolution in the diverse insect order Trichoptera. GigaScience **11**: giac011. doi:10.1093/gigascience/giac011.

Jeffery, N.W., Ellis, E.A., Oakley, T.H., and Gregory, T.R. 2017. The Genome Sizes of Ostracod Crustaceans Correlate with Body Size and Evolutionary History, but not Environment. J. Hered. **108**(6): 701–706. doi:10.1093/jhered/esx055. PMID: 28595313.

Kapitonov, V.V., and Jurka, J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. **9**(5): 411–412; author reply 414. doi:10.1038/nrg2165-c1. PMID: 18421312.

Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica **115**(1): 49–63. doi:10.1023/A:1016072014259. PMID: 12188048.

Kim, B.Y., Wang, J.R., Miller, D.E., Barmina, O., Delaney, E., Thompson, A., et al. 2021. Highly contiguous assemblies of 101 drosophilid genomes. eLife **10**: e66405. eLife Sciences Publications, Ltd. doi:10.7554/eLife.66405. PMID: 34279216.

Kraaijeveld, K. 2010. Genome size and species diversification. Evol. Biol. **37**(4): 227–233. doi:10.1007/s11692-010-9093-4. PMID: 22140283.

Lawrie, D.S., Messer, P.W., Hershberg, R., and Petrov, D.A. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet. **9**(5): e1003527. doi:10.1371/journal.pgen.1003527. PMID: 23737754.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. **38**(10): 4647–4654. doi:10.1093/molbev/msab199. PMID: 34320186.

McGowan, J. 2020, December 14. jamiemcg/BUSCO_phylogenomics: BUSCO v4. Zenodo. doi:10.5281/zenodo.4320788.

Melo, E.S., and Wallau, G.L. 2020. Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. PLoS Genet. **16**(11): e1008946. Public Library of Science. doi:10.1371/journal.pgen.1008946. PMID: 33253164.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. 2007. Which transposable elements are active in the human genome? Trends Genet. **23**(4): 183–191. doi:10.1016/j.tig.2007.02.006. PMID: 17331616.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37**(5): 1530–1534. doi:10.1093/molbev/msaa015. PMID: 32011700.

Molina-Venegas, R., and Rodríguez, M.A. 2017. Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? BMC Evol. Biol. **17**: 53. doi:10.1186%2Fs12862-017-0898-y. PMID: 28201989.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. 2012. How to measure and test phylogenetic signal. Methods Ecol. Evol. **3**(4): 743–756. doi:10.1111/j.2041-210X.2012.00196.x.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science **316**(5832): 1718–1723. doi:10.1126/science.1138878.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., and Pearse, W. 2018. caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 1.0.1. Available from https://CRAN.R-project.org/package=caper>

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**(3): 442–453. doi:10.1006/mpev.2001.1053.

O'Grady, P.M., and DeSalle, R. 2018. Phylogeny of the Genus *Drosophila*. Genetics **209**(1): 1–25. doi:10.1534/genetics.117.300583.

Palazzo, A., Escuder, E., D'Addabbo, P., Lovero, D., and Marsano, R.M. 2021. A genomic survey of Tc1-mariner transposons in nematodes suggests extensive horizontal transposon transfer events. Mol. Phylogenet. Evol. **158**: 107090. doi:10.1016/j.ympev.2021.107090.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Rebollo, R., Horard, B., Hubert, B., and Vieira, C. 2010. Jumping genes and epigenetics: towards new species. Gene **454**(1): 1–7. doi:10.1016/j.gene.2010.01.003.

Ren, L., Huang, W., Cannon, E.K.S., Bertioli, D.J., and Cannon, S.B. 2018. A mechanism for genome size reduction following genomic rearrangements. Front. Genet. **9**. Available from https://www.frontiersin.org/articles/10.3389/fgene.2018.00454 [accessed 11 August 2022]. doi:10.3389/fgene.2018.00454.

Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. **3**(2): 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

Ritchie, M.G., and Gleason, J.M. 1995. Rapid evolution of courtship song pattern in *Drosophila willistoni* sibling species. J. Evol. Biol. **8**(4): 463–479. doi:10.1046/j.1420-9101.1995.8040463.x.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F.J. 2000. Evidence for a high ancestral GC content in *Drosophila*. Mol. Biol. Evol. **17**(11): 1710–1717. doi:10.1093/oxfordjournals.molbev.a026269.

Robe, L.J., Cordeiro, J., Loreto, E.L.S., and Valente, V.L.S. 2010. Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila willistoni* subgroup (Diptera: Drosophilidae). Genetica **138**(6): 601–617. doi:10.1007/s10709-009-9432-5.

Rohde, C., and Valente, V.L.S. 2012. Three decades of studies on chromosomal polymorphism of *Drosophila willistoni* and description of fifty different rearrangements. Genet. Mol. Biol. **35**: 966–979. Sociedade Brasileira de Genética. doi:10.1590/S1415-47572012000600012.

Schaack, S., Gilbert, C., and Feschotte, C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. **25**(9): 537–546. doi:10.1016/j.tree.2010.06.001.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science **326**(5956): 1112–1115. doi:10.1126/science.1178534.

Seetharam, A.S., and Stuart, G.W. 2013. Whole genome phylogeny for 21 *Drosophila* species using predicted 2b-RAD fragments. PeerJ **1**: e226. doi:10.7717/peerj.226.

Sessegolo, C., Burlet, N., and Haudry, A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol. Lett. **12**(8): 20160407. doi:10.1098/rsbl.2016.0407.

Smit, A., Hubley, R., and Green, P. 2022. RepeatMasker Open-4.0. Available from http://www.repeatmasker.org.

Smith, S.A., and O'Meara, B.C. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. **28**(20): 2689–2690. doi:10.1093/bioinformatics/bts492.

Stelzer, C.-P., Riss, S., and Stadler, P. 2011. Genome size evolution at the speciation level: the cryptic species complex *Brachionus plicatilis* (Rotifera). BMC Evol. Biol. **11**(1): 90. doi:10.1186/1471-2148-11-90.

Su, W., Sharma, S.P., and Peterson, T. 2018. Evolutionary Impacts of Alternative Transposition. *In Origin and Evolution of Biodiversity*. *Edited by* P. Pontarotti. Springer International Publishing, Cham. pp. 113–130. doi:10.1007/978-3-319-95954-2_7.

Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., et al. 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr. Biol. **32**: 111–123. doi:10.1016/j.cub.2021.10.052.

Tarrío, R., Rodríguez-Trelles, F., and Ayala, F.J. 2000. Tree rooting with outgroups when they differ in nucleotide composition from the ingroup: the *Drosophila saltans* and *Drosophila willistoni* groups, a case study. Mol. Phylogenet. Evol. **16**(3): 344–349. doi:10.1006/mpev.2000.0813.

Tarrío, R., Rodríguez-Trelles, F., and Ayala, F.J. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. **18**(8): 1464–1473. doi:10.1093/oxfordjournals.molbev.a003932.

Vicario, S., Moriyama, E.N., and Powell, J.R. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol. Biol. **7**: 226. doi:10.1186/1471-2148-7-226.

Waltari, E., and Edwards, S.V. 2002. Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. Am. Nat. **160**(5): 539–552. doi:10.1086/342079.

Wells, J.N., and Feschotte, C. 2020. A field guide to eukaryotic transposable elements. Annu. Rev. Genet. **54**: 539–561. doi:10.1146/annurev-genet-040620-022145.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**(12): 973–982. doi:10.1038/nrg2165.

Winge, H. 1965. Interspecific hybridisation between the six cryptic species of *Drosophila willistoni* group. Heredity **20**(1): 9–19. Nature Publishing Group. doi:10.1038/hdy.1965.2.

Winge, H., and Cordeiro, A.R. 1963. Experimental hybrids between *Drosophila willistoni* sturtevant and *Drosophila paulistorum* Dobzhansky and Pavan from southern marginal populations. Heredity **18**(2): 215–222. Nature Publishing Group. doi:10.1038/hdy.1963.23.

Yassin, A. 2013. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Syst. Entomol. **38**: 349–364. doi:10.1111/j.1365-3113.2012.00665.x.

Zanini, R., Müller, M.J., Vieira, G.C., Valiati, V.H., Deprá, M., and Valente, V.L.S. 2018. Combining morphology and molecular data to improve *Drosophila paulistorum* (Diptera, Drosophilidae) taxonomic status. Fly **12**(2): 81–94. Taylor & Francis. doi:10.1080/19336934.2018.1429859.

Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., and Gilbert, C. 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat. Commun. **11**(1): 1362. doi:10.1038/s41467-020-15149-4.

Supplementary material

Supplementary Material 1 – Figures

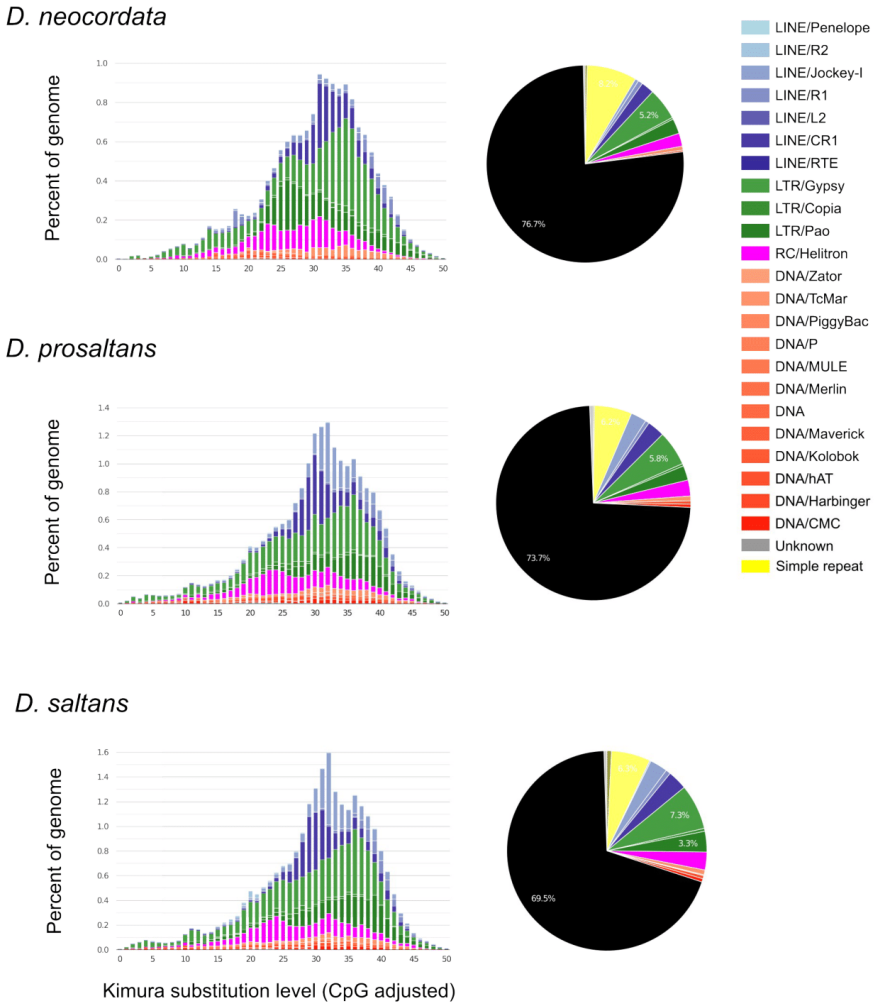


Figure S1. Interspersed repeat landscape and fractions of repetitive elements for the species belonging to the *saltans* group used as outgroup.

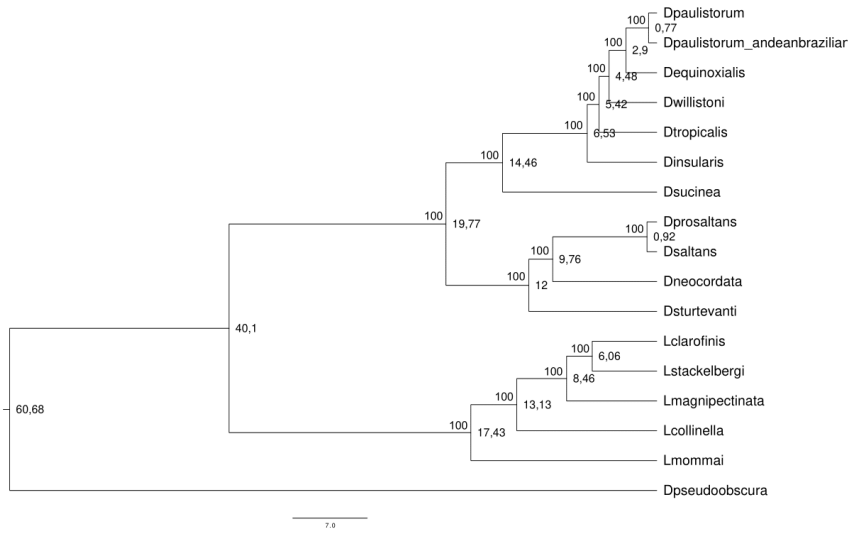


Figure S2. Maximum likelihood tree based on the supermatrix data. Numbers next to each node indicate its age, in million years; numbers above each node indicate ultrafast bootstrap support.

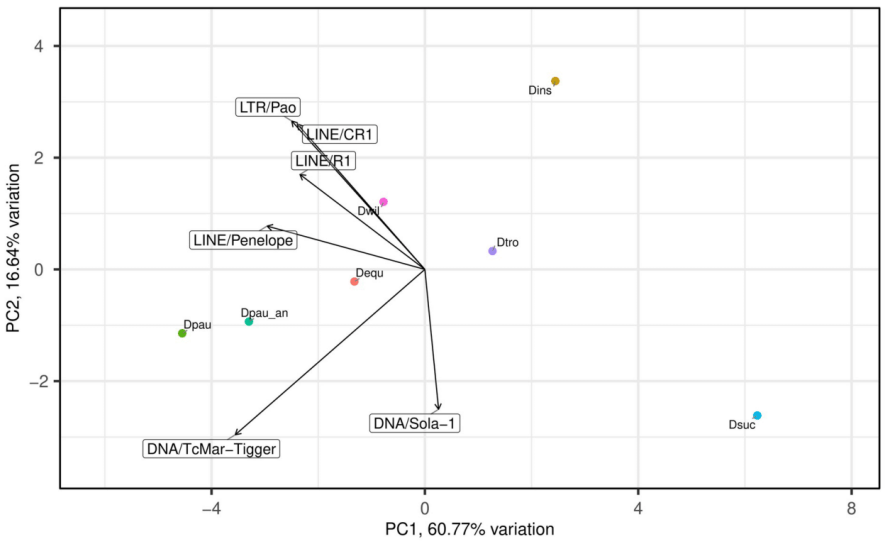


Figure S3. Principal Component Analysis (PCA) on of copy number of superfamilies of transposable elements found in the analyzed genomes. Arrows indicate superfamilies that most contribute to the variation in each principal component. Axes X and Y represent, respectively, principal components 1 and 2.

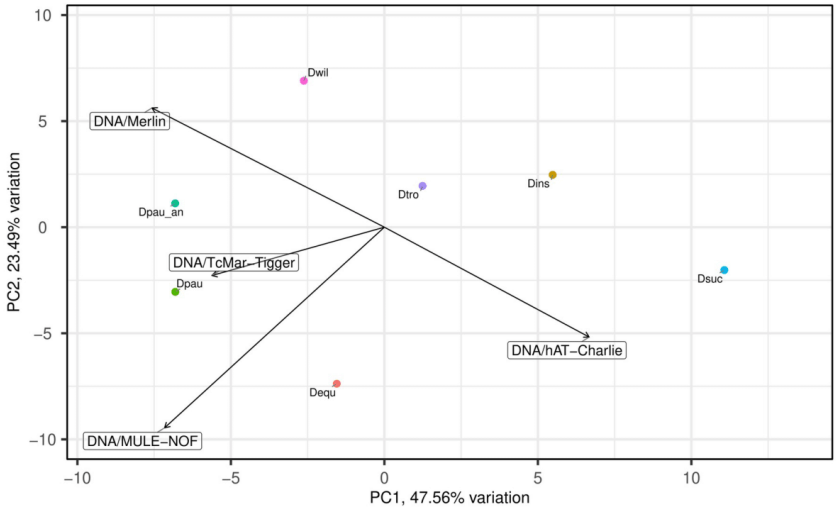


Figure S4. Principal Component Analysis (PCA) on total length (in base pairs) of superfamilies of transposable elements found in the analyzed genomes. Arrows indicate superfamilies that most contribute to the variation in each principal component. Axes X and Y represent, respectively, principal components 1 and 2.

Table S1. Total copy number of each superfamily annotated by RepeatMasker.

A	Dequ	Dins	Dpau	Dpau_an	Dsuc	Dtro	Dwil
DNA/CMC-Chapaev-3	285	336	314	255	41	279	230
DNA/CMC-Transib	380	274	535	450	282	278	320
DNA/hAT	312	231	481	348	44	98	261
DNA/hAT-Ac	526	481	806	683	393	329	565
DNA/hAT-Blackjack	50	17	43	43	4	6	56
DNA/hAT-Charlie	3	2	0	0	1	1	0
DNA/hAT-hobo	556	281	715	658	274	255	547
DNA/hAT-Pegasus	79	53	62	60	62	51	86
DNA/hAT-Tip100	37	24	349	246	35	41	56
DNA/Kolobok-Hydra	29	5	300	255	105	64	25
DNA/Maverick	14	27	36	21	14	15	18
DNA/Merlin	0	0	5	3	0	1	2
DNA/MULE-NOF	3	0	2	1	0	0	0
DNA/P	713	445	769	585	255	613	1064
DNA/PIF-Harbinger	3393	5287	7825	5496	1277	2664	3470
DNA/PiggyBac	48	49	213	102	22	25	97
DNA/RC	10303	7053	19800	15910	5135	8717	12154
DNA/Sola-1	24	0	2	0	4	2	0
DNA/TcMar	8	4	15	13	1	0	0
DNA/TcMar-Mariner	147	92	193	170	113	50	87
DNA/TcMar-Tc1	261	294	364	510	237	244	329
DNA/TcMar-Tigger	206	4	267	261	11	50	108
DNA/Zator	10	0	8	7	0	3	4
LINE/CR1	4504	4087	6960	5557	287	3733	4227
LINE/I	181	229	227	214	300	172	368
LINE/I-Jockey	1640	1527	2585	2079	887	1730	2062
LINE/Penelope	929	358	1936	1471	59	371	569
LINE/R1	994	865	2277	1723	104	1069	963
LINE/R1-LOA	2	4	2	2	2	1	3
LINE/R2	115	18	188	63	20	57	99
LINE/RTE-BovB	0	0	0	1	0	0	0
LTR/Copia	679	419	1376	1002	122	641	834
LTR/Gypsy	10388	9009	20471	15378	1848	10981	14122
LTR/Pao	5821	5470	10169	7646	362	4927	6108

Table S2. Total length (in base pairs) of each superfamily annotated by RepeatMasker.

A	Dequ	Dins	Dneb	Dpau	Dpau_an	Dsuc	Dtro	Dwil
DNA/CMC-Chapaev-3	184576	173499	6315	140573	110063	5227	181646	107803
DNA/CMC-Transib	309683	179180	179747	451601	391412	167891	219762	240605
DNA/hAT	102580	37590	4841	111940	87449	5658	15717	87759
DNA/hAT-Ac	141882	106712	91688	199471	142922	84189	84610	119052
DNA/hAT-Blackjack	24234	10208	1745	17857	17593	1836	5306	41597
DNA/hAT-Charlie	263	174	0	0	0	96	80	0
DNA/hAT-hobo	268214	140059	89315	383897	386547	93862	137467	303948
DNA/hAT-Pegasus	20562	8394	14545	19581	15034	9289	17542	18454
DNA/hAT-Tip100	9474	4125	9628	72726	49042	9125	10893	15605
DNA/Kolobok-Hydra	10233	2625	433	13380	16756	290	7902	9070
DNA/Maverick	1217	3001	1421	3217	1798	1277	1377	1572
DNA/Merlin	0	0	0	400	240	0	80	160
DNA/MULE-NOF	327	0	0	208	196	0	0	0
DNA/P	195846	90795	42368	180878	139172	45909	212704	413181
DNA/PIF-Harbinger	746571	1105659	224779	1592025	1205234	164761	598711	767427
DNA/PiggyBac	37514	15366	4896	103977	45204	4025	12074	45686
DNA/RC	4501946	1976448	1780780	8655594	7161161	1455664	3834934	5478152
DNA/Sola-1	5181	0	203	408	0	730	355	0
DNA/TcMar	1100	443	0	1994	1879	86	0	0
DNA/TcMar-Mariner	73285	32333	37473	103512	89758	35325	19993	36902
DNA/TcMar-Tc1	138186	144462	88351	176458	182413	91555	149972	183102
DNA/TcMar-Tigger	33784	365	1482	44021	42593	1373	7213	17553
DNA/Zator	1318	0	0	822	735	0	313	658
LINE/CR1	8256477	7240753	163534	14156288	10900002	156721	6700346	8338117
LINE/I	109814	99971	98042	123331	105306	80767	149223	129073
LINE/I-Jockey	2389018	2086845	126009	4136672	3220618	121283	2481528	3097962
LINE/Penelope	801374	265900	17917	1771974	1351591	13134	346849	408275
LINE/R1	2139232	1584359	73210	5337296	3719581	196634	2057180	2022217
LINE/R1-LOA	174	472	223	200	202	337	85	294
LINE/R2	129142	16657	21617	301461	54275	26246	57697	71570
LINE/RTE-BovB	0	0	0	0	84	0	0	0
LTR/Copia	1243309	626289	324586	2497798	1687249	243235	1308595	1612741
LTR/Gypsy	17893091	14499260	1729889	35399217	24895857	1578674	20779129	26916855
LTR/Pao	10343169	8411366	195591	19902027	14186497	170090	9162726	12627076

Table S3. Phylogenetic Generalized Least Squares (PGLS) statistical results for the first model.

	Value	Std error	t-value	p-value
Intercept	5.683034	0.063771	89.1169	3.372e-09
REF	0.204039	0.038171	5.3454	0.003076

Model: GS ~FRE; FRE: fraction of repetitive elements; GS: genome size.
Residual standard error: 0.009313 on 5 degrees of freedom
Multiple R-squared: 0.8511, Adjusted R-squared: 0.8213
F-statistic: 28.57 on 1 and 5 DF, p-value: 0.003076

Table S4. Phylogenetic Generalized Least Squares (PGLS) statistical results for the second model.

	Value	Std error	t-value	p-value
Intercept	-1.473550	2.849454	-0.5171	0.6271
GS	0.081094	0.519568	0.1561	0.8821

Model: FRE ~ GS; FRE: fraction of repetitive elements; GS: genome size.
Residual standard error: 5.978e-05 on 5 degrees of freedom
Multiple R-squared: 0.004848, Adjusted R-squared: -0.1942
F-statistic: 0.02436 on 1 and 5 DF, p-value: 0.8821

CAPÍTULO IV

DNA and LTR transposable elements are the main type of repetitive elements responsible for the intraspecific genome size variation in *Drosophila willistoni* (Diptera, Drosophilidae)

Henrique R. M. Antoniolli

Sebastián Pita

Beatriz Goñi

Maríndia Deprá

Vera L. S. Valente

Manuscrito a ser submetido ao periódico *Molecular Genetics and Genomics* (ISSN: 1617-4623)

CAPÍTULO V

Transposable element-mediated chromosomal inversions in the (un)stable genome of *Drosophila willistoni* (Diptera: Drosophilidae)

Henrique R.M. Antoniolli

Sebastián Pita

Beatriz Goñi

Maríndia Deprá

Vera L.S. Valente

Manuscrito a ser submetido ao periódico *Genome Biology and Evolution* (ISSN: 1759-6653)

CAPÍTULO VI

Horizontal transfer and the widespread presence of *Galileo* transposons in Drosophilidae (Insecta: Diptera)

Henrique R.M. Antoniolli

Sebastián Pita

Maríndia Deprá

Vera L.S. Valente

Manuscrito publicado no periódico *Genetics and Molecular Biology* (ISSN: 1480-3321)

doi: 10.1590/1678-4685-GMB-2023-0143

Genetics and Molecular Biology, 46, 3(suppl 1), e20230143 (2023)

DOI: https://doi.org/10.1590/1678-4685-GMB-2023-0143

Research Article
60 years of the PPGBM UFRGS – Special Issue

Horizontal transfer and the widespread presence of *Galileo* transposons in Drosophilidae (Insecta: Diptera)

Henrique R.M. Antoniolli[1], Sebastián Pita[2], Maríndia Deprá[1] and Vera L.S. Valente[1]

[1]*Universidade Federal do Rio Grande do Sul (UFRGS), Laboratório de Drosophila, Programa de Pós-Graduação em Genética e Biologia Molecular, Porto Alegre, RS, Brazil.*
[2]*Universidad de la República (UdelaR), Facultad de Ciencias, Sección Genética Evolutiva, Montevideo, Uruguay.*

Abstract

Galileo is a transposon notoriously involved with inversions in *Drosophila buzzatii* by ectopic recombination. Although widespread in *Drosophila*, little is known about this transposon in other lineages of Drosophilidae. Here, the abundance of the canonical *Galileo* and its evolutionary history in Drosophilidae genomes was estimated and reconstructed across genera within its two subfamilies. Sequences of this transposon were masked in these genomes and their transposase sequences were recovered using BLASTn. Phylogenetic analyses were employed to reconstruct their evolutionary history and compare it to that of host genomes. *Galileo* was found in nearly all 163 species, however, only 37 harbored nearly complete transposase sequences. In the remaining, *Galileo* was found highly fragmented. Copies from related species were clustered, however horizontal transfer events were detected between the *melanogaster* and *montium* groups of *Drosophila*, and between the latter and the *Lordiphosa* genus. The similarity of sequences found in the *virilis* and *willistoni* groups of *Drosophila* was found to be a consequence of lineage sorting. Therefore, the evolution of *Galileo* is primarily marked by vertical transmission and long-term inactivation, mainly through the deletion of open reading frames. The latter has the potential to lead copies of this transposon to become miniature inverted-repeat transposable elements.

Keywords: DNA transposon, MITEs, *P* superfamily.

Received: May 08, 2023; Accepted: January 30, 2024.

Introduction

Transposable elements (TEs) belong to the repetitive fraction of genomes, and are linear sequences of DNA that have the ability to move within or between genomes (Wells and Feschotte, 2020). Classifications divide these sequences firstly into two classes, based on the intermediate molecule in their transposition process (Finnegan, 1989). Class I is composed of retrotransposons as their mobilization involves the synthesis of an RNA molecule, which are retrotranscribed into DNA and then inserted elsewhere in the genome (Wicker *et al.*, 2007). On the other hand, the majority of Class II elements – or DNA transposons – are directly excised by their transposase (TPase), and then reinserted in another site in the genome (Wicker *et al.*, 2007).

In addition, TEs can be either autonomous or nonautonomous (Wicker *et al.*, 2007). The first are those that present their structures preserved, encoding all necessary enzymes to be transposed. The latter comprise defective TEs that no longer encode nor produce their own proteins, and move only if recognized by the enzymes of a closely related autonomous TE; such as the Miniature Inverted-repeat Transposable Elements (MITEs). MITEs are non-autonomous TEs, derived from autonomous Class II transposons, and present a few structural characteristics: (i) small size, ranging from 50 to 500 base pairs (bp); (ii) AT-rich sequences; and (iii) a lack of a functional TPase (Deprá *et al.*, 2012; Fattash *et al.*, 2013).

Transposable elements are often referred to as "parasites" (Colonna Romano and Fanti, 2022), given their ability to invade new genomes and increase their copy number (Loreto *et al.*, 2008). Horizontal transposon transfer (HTT) is the phenomenon in which a given TE "jumps" to the genome of a non-closely related species, i.e., sexually isolated organisms (Panaud, 2016). The role of HTT in shaping diversity as an endogenous source of evolution is widely recognized (Pace *et al.*, 2008; Gilbert and Feschotte, 2018; Carvalho *et al.*, 2023), and its frequency is much higher than previously thought (Schaack *et al.*, 2010; Panaud, 2016; Peccoud *et al.*, 2017; Melo and Wallau, 2020).

In this sense, several evolutionary events have been proposed as a direct consequence of TEs mobilization and/or recombination. For instance, in several taxa the variation and evolution of genome size are directly related to the amplification or contraction in TEs copy number (Canapa *et al.*, 2015; Antoniolli *et al.*, 2023). Nucleotide polymorphisms are also frequently produced after transposition events (Bourque *et al.*, 2018). Transposable elements are also known to be related to changes in gene expression, either by silencing or enhancing them (Finnegan, 1989), and chromosomal rearrangements – i.e., deletions, duplications, translocations and inversions by ectopic recombination (Kidwell and Lisch,

Send correspondence to Maríndia Deprá. Universidade Federal do Rio Grande do Sul (UFRGS), Laboratório de Drosophila, Avenida Bento Gonçalves, 9500, Agronomia, 91509-900, Porto Alegre, RS, Brazil. E-mail: marindiadepra@gmail.com.

1997). In the latter, distant loci in a genome carry highly similar TE copies, which allows homologous recombination to occur (see review in Bourque *et al.*, 2018), thus resulting in a drastic modification in the chromosome architecture (Ren *et al.*, 2018). Documented cases of a TE as a mediator of ectopic recombination include the families of retrotransposons *Bel-Pao*, *Doc*, *I* element and *roo*, as well as the transposons *foldback*, *Galileo* and *hobo* (Lim and Simmons, 1994; Delprat *et al.*, 2009).

Galileo is a family of Class II transposons, and encodes its own TPase flanked by terminal inverted repeats (TIRs). Initially described as a *foldback*-like element, its TIRs and THAP domains exhibit similarities with those of the *P* element, leading to the classification of *Galileo* within the *P* superfamily (Marzo *et al.*, 2008). However, unlike the *P* element, *Galileo* does not present introns (Marzo *et al.*, 2008). *Galileo* was discovered by Cáceres *et al.* (1999) due to its association with the breakpoints of the *2j* inversion in wild specimens of *Drosophila buzzatii*. In fact, *Galileo* is the only TE known to induce chromosomal rearrangements in natural populations of *Drosophila* (Marzo *et al.*, 2008), as most others have been observed in laboratory populations (Lim and Simmons, 1994). Besides the *2j* inversion, *Galileo* was involved with two other rearrangements described in *D. buzzatii* (Casals *et al.*, 2003; Delprat *et al.*, 2009). This makes this transposon as one of the most well-documented examples of a natural TE-induced chromosomal rearrangement.

Studies have shown the widespread presence of *Galileo* in the *Drosophila* genus (Marzo *et al.*, 2008; Acurio, 2015). The main focus of the present study was to characterize the evolutionary history of the *Galileo* family and evaluate its main transmission mode in Drosophilidae. This transposon was masked in genome assemblies of 163 species available at online databases, and TPase sequences found were employed for reconstructing a phylogeny and testing putative cases of HTT.

Material and Methods

Masking *Galileo* in the genome assemblies

Representative genome assemblies of 163 Drosophilidae species (see details on taxonomy and accession numbers in Table S1) were retrieved from GenBank (NCBI) with a Python package written by Blin (2021). These species belong to the *Chymomyza*, *Drosophila*, *Lordiphosa*, *Scaptodrosophila*, *Scaptomyza*, and *Zaprionus* genera of the Drosophilinae subfamily; and *Leucophenga* and *Phortica* of Steganinae subfamily (Table S1). BUSCO v.5 (Manni *et al.*, 2021) was employed to assess the completeness of each assembly with the Diptera orthologous database.

The nucleotide sequence of seven *Galileo* copies characterized by Marzo *et al.* (2008) in *D. ananassae* (Dana*Galileo* – BK006363), *D. buzzatii* (Dbuz*Galileo* – EU334682 and EU334685), *D. mojavensis* (Dmoj*Galileo* – BK006357), *D. persimilis* (Dper*Galileo* – BK006361), *D. virilis* (Dvir*Galileo* – BK006359) and *D. willistoni* (Dwil*Galileo* – BK006360) were downloaded from GenBank, and used as queries in our workflow. Firstly, the queries were input as the repeat library in RepeatMasker (Smit *et al.*, 2023) for masking canonical *Galileo* sequences in each genome assembly. The script 'One code to find them all' (Bailly-Bechet *et al.*, 2014) was then employed to parse the output, recovering the nucleotide sequence of each identified copy in an assembly with at least 80% identity to its best query and a minimum length of 80 base pairs.

Phylogenetic analysis of *Galileo* potentially autonomous copies

The complete nucleotide sequence encoding the transposase (TPase) of six copies (Dana*Galileo*, Dbuz*Galileo*, Dmoj*Galileo*, Dper*Galileo*, Dvir*Galileo*, and Dwil*Galileo*) served as queries for local BLASTn searches in each FASTA file containing the *Galileo* copies of each genome. Hits with at least 80% identity and coverage of at least 70% for any of the queries were used in downstream analyses. Additionally, a *P* element from the genome of *Drosophila buzzatii* (GenBank accession No. KC690135) and two copies of the *1360* element (GenBank accession Nos. AF533772 and AY138841) were included in the nucleotide matrix as outgroups. This matrix was aligned with MACSE v2 (Ranwez *et al.*, 2018) in two steps: (i) using the option *alignSequences*, which aligns nucleotide sequences based on their underlying codon structure, accounting for frameshifts and stop codons; (ii) the resulting alignment was edited with the option *exportAlignment*, replacing codons containing frameshifts and internal stop codons with "N" (e.g., TG! was replaced by NNN). The codon alignment was then processed with Gblocks (Castresana, 2000) to remove poorly aligned regions, allowing the presence of gaps.

The final codon alignment was translated to amino acids and used for a Bayesian phylogenetic inference (BI) analysis, performed in MrBayes 3.2.7 (Ronquist *et al.*, 2012). The majority-rule consensus tree was built under the best amino acid substitution model, as estimated by ModelTest-NG (Darriba *et al.*, 2020). Metropolis-coupled Markov chain Monte Carlo (MCMCMC) analysis was run with two parallel runs with four chains each for 1,000,000 generations, sampling every 100. Convergence was reached when the average standard deviation of split frequencies was below 1%. A burn-in of 25% was applied to the sampled trees before obtaining the consensus tree. The tree was visualized and edited in FigTree (Rambaut, 2018).

Analysis of abundance and repeat profile

Forward short-reads of high-throughput whole genome sequencing were downloaded from the Sequence Read Archive of NCBI (see SRA accession No. in Table S1) for those species with positive hits for the TPase queries. These were submitted to the RepeatProfiler pipeline (Negm *et al.*, 2021), an analysis in which sequencing reads are mapped against queries to build coverage graphs, allowing to infer which regions of a given query have a higher or lower abundance.

Quality trim was performed with fastp (Chen *et al.*, 2018), when reads had their adaptor removed while keeping only reads with no N base. The total reads were downsampled to 3 million, achieving near 1x coverage for all genomes (assuming a genome size mean of 200 megabases for species of the family Drosophilidae). In addition, five single-copy

genes were randomly selected in the Diptera orthologous genes dataset of BUSCO 5 (Manni *et al.*, 2021) to normalize the results (Table S2). The six complete copies of *Galileo* used in BLASTn searches were used as queries (Dbuz*Galileo* – EU334685 was excluded because it was shorter than Dbuz\ *Galileo* – EU334682). RepeatProfiler (Negm *et al.*, 2021) was executed with default parameters.

Inference of HTT events

Possible cases of HTT were determined based on incongruences between the phylogeny of host genomes and the phylogeny of *Galileo*. Validation of such cases was performed with the *vhica* R package (Wallau *et al.*, 2015), implemented on the HTT-DB platform (Dotto *et al.*, 2015). This method relies on discrepancies in the evolutionary rates of synonymous positions (dS), which considers codon usage bias (CUB), between nuclear genes (vertically transmitted) and transposable elements (TEs). Wallau *et al.* (2015) demonstrated that dS and CUB are correlated, and low values for both are indicative of inconsistencies with vertical transmission.

Sequences of single-copy orthologous genes were searched in the assemblies with positive hits of *Galileo* using BUSCO 5 (Manni *et al.*, 2021) and the Diptera orthologous database. Nucleotide sequences of 30 randomly selected genes (see Table S3) were aligned based on codons using the ClustalW algorithm (Thompson *et al.*, 1994) implemented in MEGA 11 (Tamura *et al.*, 2021). These alignments were used to compare the dS-CUB between the host nuclear genome and *Galileo* sequences. A substitution rate of 0.016 per million years (Sharp and Li, 1989) was applied to estimate the time of divergence between *Galileo* sequences.

To provide an evolutionary context for the results, a phylogenetic tree of the 37 host genomes was reconstructed using the entire set of BUSCO genes shared among them. Their amino acid sequences were aligned with MUSCLE (Edgar, 2004) and refined with trimAl (Capella-Gutiérrez *et al.*, 2009), implemented in a pipeline written by McGowan (2020). *Scaptodrosophila lebanonensis* was included in this analysis as an outgroup. Their phylogenetic relationships were reconstructed under maximum likelihood in IQ-TREE 2 (Minh *et al.*, 2020), with the best substitution model selected based on AIC scores (flags *--m* and *--merit*). Branch supports were estimated by applying 1,000 replicates of ultrafast bootstrap.

Results

Search for canonical *Galileo* copies

Sequences of *Galileo* were masked in all analyzed genomes (Table S1), except for *D. ercepeae* and *D. nannoptera* – which belong to the *melanogaster* and *nannoptera* groups, respectively. Assemblies showed satisfactory levels of completeness, with the majority having more than 90% of single-copy orthologous genes (S). The exception was eight species, with S percentages ranging from 70% to 90% (see Table S4). In the second round of searches, conducted using local BLASTn with TPases as queries, 37 species yielded positive hits after the filtering process (Table S1). The positive results in the BLASTn search were limited to species within the *Drosophila* and *Lordiphosa* genera (Drosophilinae subfamily, Drosophilini tribe). All identified TPase sequences exhibited mutations, including stop codons, coding frame shifts, or both.

Phylogenetic analysis and abundance of *Galileo* sequences

The final sizes of nucleotide and amino acid alignments were 1,035 bp and 345 amino acids, respectively. The best amino acid substitution model was JTT+G4+F, based on the Akaike Information Criterion (AIC). Every copy of *Galileo* found in the genomes was placed in the same clade as its query. Major clades exhibited strong node support (PP > 0.95), with exceptions mainly observed among intraspecific sequences.

The query Dana*Galileo* recovered three clades: the first two (yellow, Figure S1) containing sequences found in genomes of the *melanogaster* group (in which *D. ananassae* is phylogenetically placed); and the third (orange clade, Figure S1) containing sequences found in species of the *montium* group, along with *Lordiphosa collinella* and *L. stackelbergi* (pink sequences, Figure S1). Dwil*Galileo* clustered homologous sequences found in the *willistoni* group (light pink sequences, Figure S1), along with its sister *saltans* group (blue sequences, Figure S1). On the other hand, the sequences of *Galileo* found by Dvir*Galileo* (green clade, Figure S1) in species of the *virilis* group formed a sister clade (PP = 1.0) to those of the *willistoni* and *saltans* groups. Finally, Dper*Galileo* recovered *Galileo* from species belonging to the *obscura* group (red clade, Figure S1), and Dmoj*Galileo* retrieved sequences in *D. mojavensis* (purple clade, Figure S1). The abundance of *Galileo* sequences in these species, as assessed by the coverage analysis in RepeatProfiler, showed that the TPase region had lower coverage than that of TIRs in all cases (Figures 1 and S2-S7).

Inference of HTT events

Two major incongruences were found between host species (Figure 2A) and *Galileo* phylogenies. The first is the similarity of elements found in *Lordiphosa collinella* and *Lordiphosa stackelbergi* with species of the *montium* group (Figure 2B). The second incongruence (Figure 2C) is the clade formed by *virilis* (*Drosophila* subgenus) and *willistoni* plus *saltans* groups (*Sophophora* subgenus). No signals of HTT events were detected (p-value > 0.05) between the species of the *virilis* group and the *willistoni* and *saltans* groups (Figure 2D). However, HTT was detected (p-value < 0.05) between *L. collinella* and *L. stackelbergi* and species of the *montium* group. Signals were also detected between the *melanogaster* and *montium* groups, both belonging to the *Sophophora* subgenus (Figure 2E). Estimates of divergence times (Table S5) span from ~679 thousand years ago (*D. auraria* x *L. stackelbergi*) to ~6 million years ago (*D. carrolli* × *D. watanabei*).

Discussion

The 163 genomes analyzed in this study provided a broader sampling across Drosophilidae when compared to previous studies (Marzo *et al.*, 2008; Acurio, 2015), including many different taxonomic levels. We were able to search for *Galileo* in the genomes of the two subfamilies – Drosophilinae and, for the first time, Steganinae. Furthermore, our sampling included two tribes of the first (Colocasiomyini

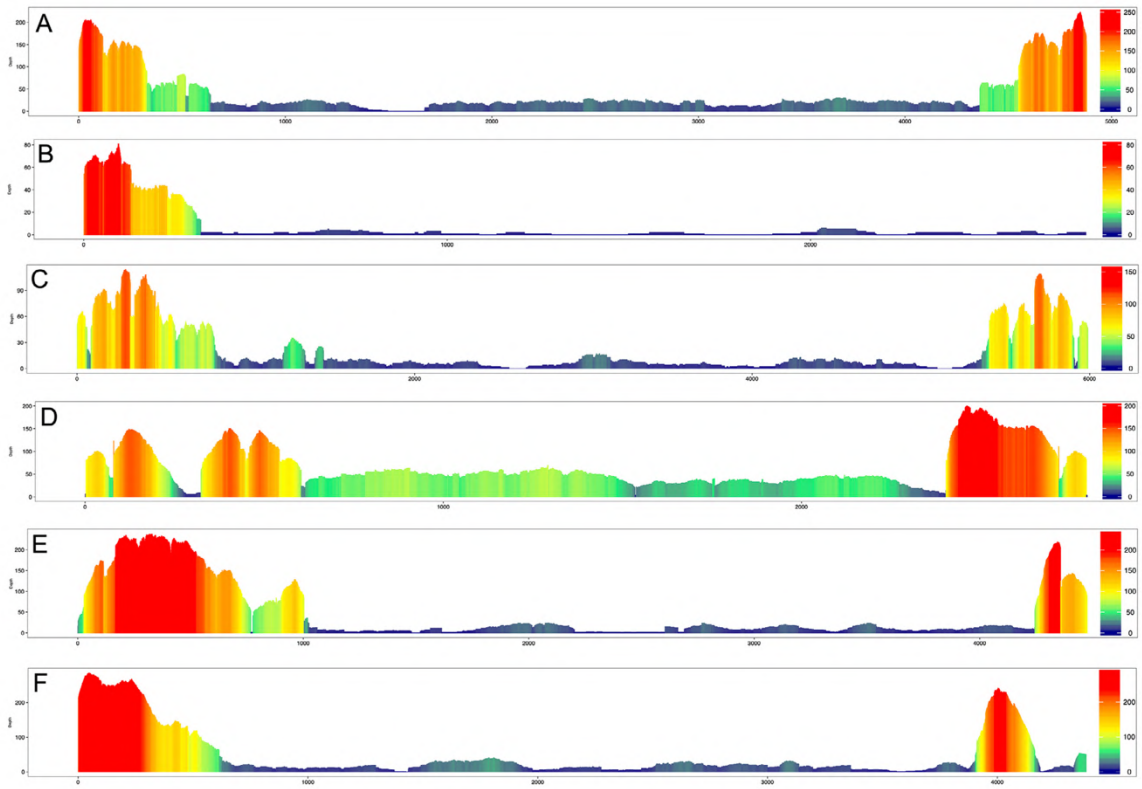


Figure 1 – Coverage graphs for six queries of *Galileo* against its corresponding species: (A) Dana*Galileo* in *Drosophila ananassae*; (B) Dbuz*Galileo* in *D. buzzatii*; (C) Dmoj*Galileo* in *D. mojavensis*; (D) Dper*Galileo* in *D. persimilis*; (E) Dvir*Galileo* in *D. virilis*; and (F) Dwil*Galileo* in *D. willistoni*. Colors correspond to the coverage scale on the right side of each graph. Axis X corresponds to base pairs positions.

and Drosophilini) and two tribes of the latter (Gitonini and Steganini). Indeed, the *Drosophila* genus is a paraphyletic lineage due to the offshoot of several genera within its phylogenetic tree (Suvorov *et al*., 2022); e.g., the *Lordiphosa* genus is placed within the *Sophophora* subgenus as a sister lineage to the Neotropical clade, which includes the *saltans* and *willistoni* groups (Figure 2D).

Galileo is fragmentally widespread in Drosophilidae

The majority of *Galileo* sequences recovered in our study consisted of fragments. Indeed, high levels of structural dynamism in *Galileo* have been described both within and between genomes, as TIRs presented variable sizes (see review in Marzo *et al*., 2008). Therefore, our results suggest that the canonical *Galileo* is widespread and abundant in the genomes of Drosophilidae, although its copies are potentially defective. Given the lack of coding for a transposase, these copies would be incapable of autonomous transposition, remaining as relics—as in the case of Miniature Inverted-repeat Transposable Elements (MITEs).

The hypothesis of classifying these fragmented copies as MITEs of *Galileo* in *D. mojavensis* was considered by Marzo *et al*. (2013a), but was discarded by those authors because the sequences were longer and had a lower copy number compared to typical MITEs. However, our analysis of normalized coverage suggested the opposite; highly amplified short segments of *Galileo* TIRs were detected (Figures 1 and S2-S7), consistent with the size of MITEs. In *D. virilis*, for example, the TPase segment of Dvir*Galileo* had a low coverage (~10X) while its TIRs had a coverage of < 200X (Figure 1E). Although strong evidence was found, further characterization is still needed to assist in the classification of these short canonical sequences as MITEs.

Interestingly, *Galileo* seems to be highly amplified in Neotropical species. Among the 15 species with the highest copy number (Figure 3; Table S1), eight are endemic to the Neotropical region: *D. mojavensis*, *D. sturtevanti*, *D. willistoni*, *D. paulistorum*, *D. navojoa*, and *D. buzzatii*, *D. tropicalis*, and *D. montana* (listed from the highest to the lowest copy number). In fact, the heterogeneity found across the Neotropical region provides innumerous distinct environments, challenging the survival of species (Miranda *et al*., 2022). Such environments also impact genomes, as expanding into new areas may relieve the epigenetic silencing or control of TEs, leading to their mobilization and amplification (Gregory, 2001; Rebollo *et al*., 2010; Antoniolli *et al*., 2023). For instance, *D. willistoni* – which harbors an exceptional diversity of *Galileo* (Gonçalves *et al*., 2014) – is distributed throughout the Neotropical region, and TEs differentially populate its genomes (Bertocchi *et al*., 2022).

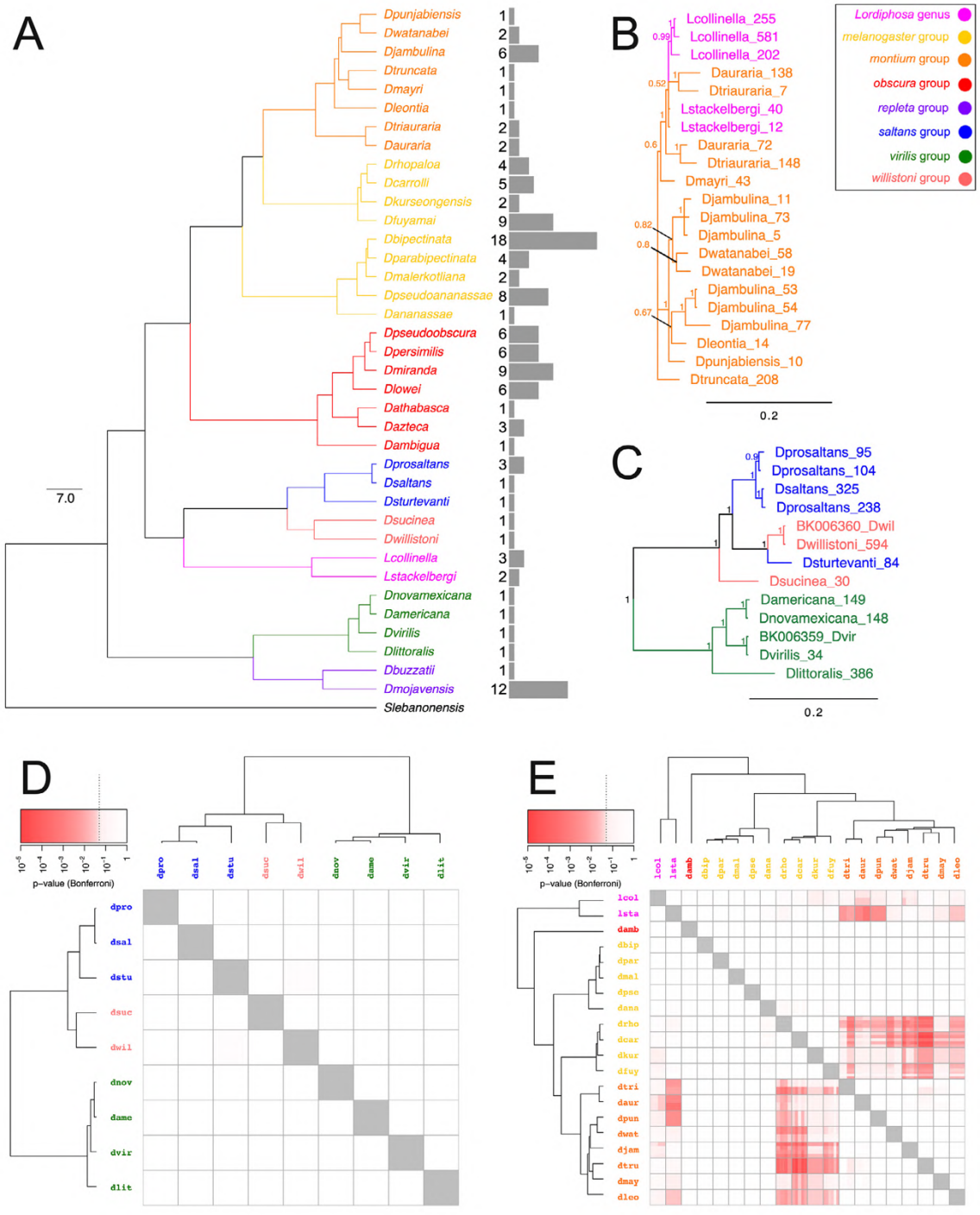


Figure 2 – (A) Ultrametric tree showing the phylogenetic relationships between species harboring nearly complete transposases, assessed through maximum likelihood. Ultrafast bootstrap (UFboot) not shown, as for all nodes UFboot = 100. (B and C) Majority-rule consensus tree showing the phylogenetic relationships between sequences of *Galileo*, (B) found in genomes of the *montium* group of *Drosophila* and species of the *Lordiphosa* genus, and (C) found in genomes of the *saltans*, *virilis* and *willistoni* groups of *Drosophila*; numbers next to each node reflect its posterior probability support. (D and E) Results of the horizontal transposon transfer (HTT) analysis in vhica, between (D) *saltans*, *virilis* and *willistoni* groups of *Drosophila*; and (E) *Lordiphosa* genus and *melanogaster* and *montium* groups of *Drosophila*. (D and E) Red squares represent statistically significant ($P < 0.05$) pairwise comparisons between sequences of *Galileo*, indicating a HTT event. Phylogenetic relationships between host genomes are shown by ultrametric trees drawn on the external sides of each graph.

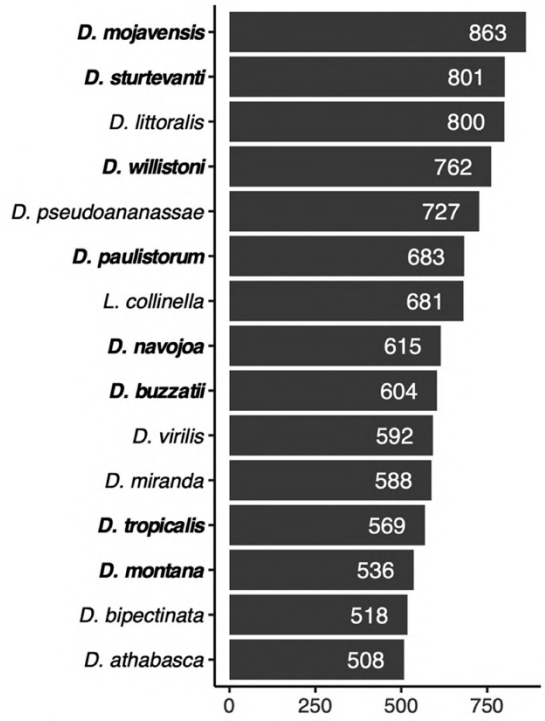


Figure 3 – Number of sequences (X axis) masked as *Galileo* elements by RepeatMasker for the top 15 species (Y axis) with the highest number of sequences. Species highlighted in bold are endemic to the Neotropical region.

Signals of HTT in the *Sophophora* subgenus

The overall congruence between the phylogeny of *Galileo* and that of its host genomes, in terms of clustering species of the same group into the same clade (Figure S1), may be explained by vertical transmission (Acurio, 2015). However, the observed incongruence involving copies found in *Lordiphosa* and species of the *montium* group (Figure 2B) was confirmed as horizontal transfer (HTT) event (Figure 2E). The *Lordiphosa* genus is actually a sister lineage to the *willistoni* group, and its MRCA with the *montium* group diverged around 40 million years ago (Mya) (Suvorov *et al.*, 2022). In this case, the oldest HTT event between them (*L. stacklbergi* × *D. punjabiensis*) is estimated to have occurred at around 2.2 Mya; much more recent than their MRCA.

Other cases of HTT involved the *melanogaster* and *montium* groups, whose MRCA diverged around 20 Mya (Suvorov *et al.*, 2022); also much older than the oldest HTT detected between them (around 6 Mya for *D. carrolli* × *D. watanabei*). The species involved with HTT events occur in sympatry, mainly in the Palearctic region of Asia (TaxoDros v1.04) – which permits a niche overlap. Addittionaly, *Galileo* exhibits a patchy distribution both in the *Lordiphosa* genus and the *melanogaster* and *montium* groups (Table S1); in this case, the TE is present in some species but absent in another closely related one(s).

Furthermore, a specific THAP binding site for the *Galileo* transposase was identified at the 3' end TIRs (Marzo *et al.*, 2013b). The sequences of *Galileo* found in these species involved in HTT cases presented highly conserved and amplified 3' TIRs (Figures 1 and S2-S7), providing further support for the plausibility of such HTT events. Nonetheless, the successful establishment of a TE in new genomes is highly dependent on its transposition rate (Le Rouzic and Capy, 2005), as it must avoid being lost in the population due to genetic drift (Blumenstiel, 2019). While *L. stackelbergi* presented a low number of sequences (49 sequences), *L. collinella* harbors more than 680 sequences (Table S1), similar to *D. buzzatii* (604 sequences), in which *Galileo* was first described. Many other cases of low copy number were also detected (Table S1), and the smallest include *D. ambigua* (10), *D. punjabiensis* (37). and *D. watanabei* (58). The process of stochastic loss of an element may explain both its patchy distribution and low copy number (Blumenstiel, 2019), as observed in *mariner*-like elements in *Drosophila* (Lohe *et al.*, 1995) and *Rex* elements in the ray-finned fish *Characidium* (Pucci *et al.*, 2018).

Lineage sorting explains the similarity between the *saltans*, *virilis* and *willistoni* groups

Marzo *et al.* (2008) described a high similarity between the copies found in the genomes of *D. virilis* and *D. willistoni*. Interestingly, the first belongs to the *Drosophila* subgenus, while the latter belongs to the *Sophophora* subgenus – their MRCA diverged around 49.9 Mya (Suvorov *et al.*, 2022). Acurio (2015) later confirmed this close relationship, identifying it along with the *guarani* and *tripunctata* groups (*Drosophila* subgenus). Our results further corroborate both studies by expanding the sample size to include *D. littoralis* and *D. novamexicana* (*virilis* group).

Interestingly, *Galileo* sequences found in each of these two groups clustered into sister clades that corresponded to their host species, with the addition of sequences from the *saltans* group in the latter. This clade (*virilis* + *saltans* + *willistoni*) was the first to split in the evolution of *Galileo* – also congruent with Marzo *et al.* (2008). These authors also proposed two explanations for the incongruence between the phylogenies of *Galileo* and its host genomes: lineage sorting with ancestral HTT (Acurio, 2015); or horizontal transfer itself. As no signal of HTT was detected between or within these three species groups (Figure 2A), lineage sorting is a plausible explanation (Cummings, 1994). In this case, the transposon is vertically transmitted, but its copies coalesce prior to the split between the host species (Tenaillon *et al.*, 2010) or are differentially lost along the branches of the species tree (Marzo *et al.*, 2008).

Conclusions

The evolutionary history of *Galileo* in Drosophilidae is marked mostly by vertical and possibly ancient horizontal transmissions, as identified by Acurio (2015), with stochastic loss through genetic drift occurring while species diverged. In addition, its high fragmentation level is compatible with the characteristics of MITEs, although a thorough characterization is still needed to confirm this. *Galileo* found favorable conditions for its amplification in the heterogeneous Neotropical region, with an astounding copy number detected in Drosophilidae species inhabiting this area. Finally, considering the potential of *Galileo* to induce chromosomal rearrangements and their

evolutionary implications, the HTT described between *Lordiphosa* and the *montium* group, and between the latter and the *melanogaster* group, these results raise an intriguing question (Alfredo Ruiz, personal communication): could evolution be infectious?

Acknowledgements

The authors thank Dr. Arnaud Le Rouzic and Dr. Gabriel L. Wallau for technical assistance with the *vhica* analysis and Dr. Alfredo Ruiz for inspiring comments in early stages of this study. This study was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under scholarship No. 141319/2020-8, and research productivity grant No. 312781/2018-0.

Conflict of Interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions

HRMA conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing-original draft, writing-review and editing; SP data curation, formal analysis, methodology, software, writing-review and editing; MD conceptualization, formal analysis, methodology, supervision, writing-review and editing; VLSV conceptualization, project administration, resources, supervision, writing-review and editing.

References

Acurio AE (2015) Coevolutionary analysis of the transposon *Galileo* in the genus *Drosophila*. M. Sc. Thesis, Autonomous University of Barcelona, 219 p.

Antoniolli HRM, Deprá M and Valente VLS (2023) Patterns of genome size evolution versus fraction of repetitive elements in *statu nascendi* species: the case of the *willistoni* subgroup of *Drosophila* (Diptera, Drosophilidae). Genome 66:193–201.

Bailly-Bechet M, Haudry A and Lerat E (2014) "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. Mob DNA 5:13.

Bertocchi NA, Oliveira TD, Deprá M, Goñi B and Valente VLS (2022) Interpopulation variation of transposable elements of the *hAT* superfamily in *Drosophila willistoni* (Diptera: Drosophilidae): *in-situ* approach. Genet Mol Biol 45:e20210287.

Blumenstiel JP (2019) Birth, school, work, death, and resurrection: The life stages and dynamics of transposable element proliferation. Genes 10:336.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL and Feschotte C (2018) Ten things you should know about transposable elements. Genome Biol 19:199.

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

Canapa A, Barucca M, Biscotti MA, Forconi M and Olmo E (2015) Transposons, genome size, and evolutionary insights in animals. Cytogenet Genome Res 147:217–239.

Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carvalho TL, Cordeiro J, Vizentin-Bugoni J, Fonseca PM, Loreto ELS and Robe LJ (2023) Horizontal transposon transfer and their ecological drivers: The case of flower-breeding Drosophila. Genome Biol Evol 15:evad068.

Casals F, Cáceres M and 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.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.

Chen S, Zhou Y, Chen Y and Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890.

Cummings MP (1994) Transmission patterns of eukaryotic transposable elements: Arguments for and against horizontal transfer. Trends Ecol Evol 9:141–145.

Colonna Romano N and Fanti L (2022) Transposable elements: Major players in shaping genomic and evolutionary patterns. Cells 11:1048.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B and Flouri T (2020) ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol 37:291–294.

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

Deprá M, Ludwig A, Valente VL and Loreto EL (2012) *Mar*, a MITE family of *hAT* transposons in *Drosophila*. Mob DNA 3:13.

Dotto BR, Carvalho EL, Silva AF, Duarte Silva LF, Pinto PM, Ortiz MF and Wallau GL (2015) HTT-DB: Horizontally transferred transposable elements database. Bioinformatics 31:2915–2917.

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

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P and Yang G (2013) Miniature inverted-repeat transposable elements: Discovery, distribution, and activity. Genome 56:475–486.

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

Gilbert C and Feschotte C (2018) Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. Curr Opin Genet Dev 49:15–24.

Gonçalves JW, Valiati VH, Delprat A, Valente VL and Ruiz A (2014) Structural and sequence diversity of the transposon *Galileo* in the *Drosophila willistoni* genome. BMC Genomics 15:792.

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76:65–101.

Kidwell MG and Lisch D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94:7704–7711.

Le Rouzic A and Capy P (2005) The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. Genetics 169:1033–1043.

Lim JK and Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays 16:269–275.

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995) Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. Mol Biol Evol 12:62–72.

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

Marzo M, Puig M and Ruiz A (2008) The *Foldback*-like element *Galileo* belongs to the *P* superfamily of DNA transposons and is widespread within the *Drosophila* genus. Proc Natl Acad Sci U S A 105:2957–2962.

Marzo M, Bello X, Puig M, Maside X and Ruiz A (2013a) Striking structural dynamism and nucleotide sequence variation of the transposon *Galileo* in the genome of *Drosophila mojavensis*. Mob DNA 4:6.

Marzo M, Liu D, Ruiz A and Chalmers R (2013b) Identification of multiple binding sites for the THAP domain of the *Galileo* transposase in the long terminal inverted-repeats. Gene 525:84-91.

Manni M, Berkeley MR, Seppey M, Simão FA and Zdobnov EM (2021) BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38(10):4647–4654.

Melo ES and Wallau GL (2020) Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. PLOS Genetics 16:e1008946.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534.

Miranda FR, Machado AF, Clozato CL and Silva SM (2022) Nine biomes and nine challenges for the conservation genetics of Neotropical species, the case of the vulnerable giant anteater (*Myrmecophaga tridactyla*). Biodivers Conserv 31:2515–2541.

Negm S, Greenberg A, Larracuente AM and Sproul JS (2021) RepeatProfiler: A pipeline for visualization and comparative analysis of repetitive DNA profiles. Mol Ecol Resour 21:969–981.

Pace JK, Gilbert C, Clark MS and Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A 105:17023–17028.

Peccoud J, Loiseau V, Cordaux R and Gilbert C (2017) Massive horizontal transfer of transposable elements in insects. Proc Natl Acad Sci U S A 114:4721–4726.

Panaud O (2016) Horizontal transfers of transposable elements in eukaryotes: The flying genes. C R Biol 7–8: 296–299.

Pucci MB, Nogaroto V, Moreira-Filho O and Vicari R (2018) Dispersion of transposable elements and multigene families: Microstructural variation in *Characidium* (Characiformes: Crenuchidae) genomes. Genet Mol Biol 41:585–592.

Ranwez V, Douzery EJP, Cambon C, Chantret N and Delsuc F (2018) MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol Biol Evol 35:2582–2584.

Rebollo R, Horard B, Hubert B and Vieira C (2010) Jumping genes and epigenetics: Towards new species. Gene 454:1–7.

Ren L, Huang W, Cannon EKS, Bertioli DJ and Cannon SB (2018) A mechanism for genome size reduction following genomic rearrangements. Front Genet 9:454.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA and Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.

Schaack S, Gilbert C and Feschotte C (2010) Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546.

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

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell PJ, Lang M, Courtier-Orgogozo V *et al.* (2022) Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr Biol 32:111–123.

Tamura K, Stecher G and Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38:3022–3027.

Tenaillon MI, Hollister JD and Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471–478.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Wallau GL, Capy P, Loreto E, Le-Rouzic A and Hua-Van A (2015) VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the *Mariner* Family within *Drosophila*. Mol Biol Evol 33:1094–1109.

Wells JN and Feschotte C (2020) A field guide to eukaryotic transposable elements. Annu Rev Genet 54:539–561.

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

Internet Resources

TaxoDros. The database on Taxonomy of Drosophilidae v1.04, https://www.taxodros.uzh.ch/ (accessed 10 April 2023)

Blin K (2021) NCBI Genome Downloading Scripts, https://github.com/kblin/ncbi-genome-download/ (accessed 10 November 2022)

McGowan J (2020) jamiemcg/BUSCO_phylogenomics: BUSCO v4, https://zenodo.org/records/7334954 (accessed 8 April 2023)

Rambaut A (2018) FigTree v1.4.4, https://github.com/rambaut/figtree/ (accessed 9 April 2023)

Smit A, Hubley R and Green P (2023) RepeatMasker 4.0, http://www.repeatmasker.org/ (accessed 10 November 2022)

Supplementary material

The following online material is available for this article:

Figure S1 – Phylogenetic relationships between sequences of *Galileo* found across genomes of Drosophilidae, reconstructed through Bayesian Inference.

Figure S2 – Normalized coverage graphs for Dana*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S3 – Normalized coverage graphs for Dbuz*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S4 – Normalized coverage graphs for Dmoj*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S5 – Normalized coverage graphs for Dper*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S6 – Normalized coverage graphs for Dvir*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S7 – Normalized coverage graphs for Dwil*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Table S1 – List and taxonomy of Drosophilidae species included in this study, including positive results from BLASTn searches of the *Galileo* transposase sequences and accession numbers to the genome assembly and short-read sequencing data on NCBI.

Table S2 – List of genes used to normalize the results of profile and abundance of *Galileo* across the analyzed genomes in this study.

Table S3 – List of genes used for Codon Usage Bias (CUB) comparisons in the analysis of horizontal transposon transfer (HTT) in vhica R package.

Table S4 – Statistics of assembly completeness for each analyzed genome.

Table S5 – Statistically significant results of pairwise comparisons of horizontal transposon transfer performed with vhica at HTT-DB.

Associate Editor: Loreta Brandão de Freitas

CAPÍTULO VII

1. Considerações finais e perspectivas

As espécies do grupo *willistoni* de *Drosophila* vêm sendo utilizadas desde meados do século XX como organismos modelo para a genética evolutiva. Os diferentes graus de isolamento reprodutivo e níveis taxonômicos em que essas espécies se apresentam foram, inclusive, utilizados como exemplos na formulação da síntese evolutiva moderna (Dobzhansky et al. 1977). Nesta tese, o grupo *willistoni* foi empregado para elucidar o papel dos elementos transponíveis (TEs – do Inglês, *transposable elements*) na evolução do genoma em espécies neotropicais incipientes. Diferentes aspectos desse processo evolutivo foram abordados ao longo de cinco capítulos, desde as relações filogenéticas entre as espécies alvo até mecanismos moleculares subjacentes à variação do tamanho do genoma e a rearranjos cromossômicos.

1.1. As relações filogenéticas no grupo *willistoni*

As primeiras sequências nucleotídicas de *D. bocainensis* para três genes mitocondriais e cinco genes nucleares foram apresentadas no Capítulo II e empregadas para reavaliar a parafilia do subgrupo *bocainensis* em relação ao subgrupo *willistoni*. A inclusão de *D. bocainensis* não foi capaz de recuperar a monofilia do seu subgrupo; contudo, adicionou um forte suporte à presença de duas linhagens distintas: a primeira, composta por *D. bocainensis*, *D. capricorni* e *D. sucinea*; e a segunda, composta por *D. fumipennis* e *D. nebulosa*, formando um clado irmão ao subgrupo *willistoni*.

As relações filogenéticas entre os subgrupos do grupo *willistoni*, portanto, ainda permanecem por ser esclarecidas. A inclusão do restante das espécies que compõem o subgrupo *bocainensis*, bem como as seis espécies pertencentes ao subgrupo *alagitans*, revelarão em definitivo a história evolutiva dessas espécies. Uma revisão taxonômica poderá, ainda, ser necessária para reavaliar as sinapomorfias de cada subgrupo ou linhagem.

1.2. A evolução do tamanho de genoma em espécies e populações

Nos últimos anos, o número crescente de genomas completos sequenciados pertencentes a espécies e subespécies da família Drosophilidae possibilitou a realização de estudos genômicos comparativos nesta tese. Assim, os capítulos III e IV versaram acerca do

papel de TEs na evolução do tamanho de genomas; primeiramente sob um enfoque interespecífico no subgrupo *willistoni*, e posteriormente, de uma forma complementar, visando a evolução intraespecífica em *D. willistoni*. No capítulo III, portanto, ao comparar o mobiloma de várias espécies e subespécies, descobrimos padrões de mobilização recente de TEs associados à história evolutiva desses táxons. Os resultados obtidos sugerem uma possível ligação entre a especiação em curso e o aumento observado no conteúdo de TEs e tamanho de genoma, oferecendo novas perspectivas sobre os mecanismos que impulsionam a divergência genômica. A caracterização intraespecífica do mobiloma em *D. paulistorum* – e, portanto, entre subespécies que estão ativamente em especiação – está sendo conduzida e poderá contestar ou confirmar a relação "especiação *versus* TEs *versus* tamanho do genoma" previamente descrita.

No capítulo IV, o genoma total de cinco linhagens de *D. willistoni* foi sequenciado e, junto a genomas de outras cinco linhagens disponíveis em bancos de dados, a variação intraespecífica do tamanho de genoma em *D. willistoni* foi aprofundada. Essa espécie apresenta altos níveis de polimorfismo cromossômico, sendo as elevadas proporções de TEs descritas no capítulo anterior uma das principais hipóteses para a instabilidade genômica observada nessa espécie. Os resultados revelaram diferenças no tamanho do genoma e, principalmente, na composição de TEs entre as linhagens. Os retrotransposons com LTR se apresentaram como principais promotores dessas diferenças – tanto na quantidade de cópias quanto nas taxas de mobilização estimadas pelas paisagens de elementos repetitivos. A análise filogenética entre as linhagens sugeriu, ainda, a coocorrência das subespécies *D. w. willistoni* e *D. w. winge* no sul da América do Sul. Uma análise filogeográfica ampla se faz aqui imprescindível para refutar ou corroborar essa hipótese.

1.3. Inversões cromossômicas induzidas por elementos transponíveis

Os resultados obtidos com os capítulos III e IV concluíram que os TEs desempenharam um papel central na evolução genômica do grupo *willistoni*, além de representarem um importante fator na instabilidade cromossômica em *D. willistoni*. No capítulo V, pela primeira vez, foi realizada em todo o genoma uma busca e anotação *in silico* de TEs associados a rearranjos cromossômicos nessa espécie. As análises identificaram numerosas inversões na linhagem uruguaia SG12.00 potencialmente induzidas por elementos transponíveis – principalmente envolvendo as famílias *Helitron*, *Gypsy* e *Galileo*.

A detecção de TEs ativos dentro dos pontos de quebra de inversão sugere, portanto, um reordenamento genômico contínuo mediado por essas sequências, enfatizando ainda mais seu impacto na evolução do genoma de *D. willistoni*. Análises *in silico* em outras linhagens, associadas à hibridização *in situ* em cromossomos politênicos, devem fornecer um suporte maior aos resultados obtidos e potencialmente indicar outras famílias associadas a rearranjos cromossômicos.

1.4. A história evolutiva do transposon *Galileo*

O elevado número de cópias do transposon *Galileo* encontradas no genoma de *D. willistoni* (Gonçalves et al. 2014), associado à descrição de inversões potencialmente induzidas por esse TE, motivou a sua busca e a reconstrução da sua história evolutiva em Drosophilidae no capítulo VI. Esse transposon se apresentou amplamente distribuído no genoma de drosofilídeos, particularmente em espécies neotropicais. No entanto, a maioria das cópias encontradas são constituídas por pequenos fragmentos que não codificam a enzima transposase (necessária para sua mobilização) – compatível com características que definem os MITEs (do Inglês, *miniature inverted-repeat transposable element*). Análises mais profundas e específicas para a caracterização de MITEs devem ser executadas para confirmar a hipótese de que esse TE está se tornando um MITE nos genomas de Drosophilidae.

A história evolutiva do elemento se mostrou incongruente com as relações filogenéticas de algumas de suas espécies hospedeiras, indicando possíveis eventos de transferência horizontal (HTT – do Inglês, *horizontal transposon transfer*). Análises específicas inferiram HTT entre os grupos *melanogaster* e *montium* do gênero *Drosophila*, e entre o último e o gênero *Lordiphosa*. Por outro lado, as HTT hipotetizadas pelas reconstruções filogenéticas entre os grupos *willistoni* e *virilis* de *Drosophila* revelaram ser, na verdade, resultado de uma seleção incompleta de linhagem (ou, do Inglês, *incomplete lineage sorting*). Esse resultado corrobora a hipótese de Marzo et al. (2008), onde *Galileo* foi transmitido verticalmente e suas cópias coalescem antes da separação entre as espécies hospedeiras. Uma caracterização *in silico* mais aprofundada desse elemento foi parcialmente executada em espécies e subespécies do grupo *willistoni*, bem como em diferentes linhagens de *D. willistoni*. Ainda permanecem por serem executadas, contudo, análises *in situ* para confirmar o observado computacionalmente.

REFERÊNCIAS BIBLIOGRÁFICAS

Arkhipova IR and Yushenova IA (2023) To be mobile or not: the variety of reverse transcriptases and their recruitment by host genomes. Biochemistry (Mosc) 88(11):1754-1762. doi: 10.1134/S000629792311007X.

Bächli G (2023). TaxoDros: the database on taxonomy of Drosophilidae. Available at https://www.taxodros.uzh.ch/.

Baião GC, Schneider DI, Miller WJ and Klasson L (2023) Multiple introgressions shape mitochondrial evolutionary history in *Drosophila paulistorum* and the *Drosophila willistoni* group. Mol Phylogenet Evol 180:107683.

Balachandran P, Walawalkar IA, Flores JI, Dayton JN, Audano PA and Beck CR (2022) Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun 13(1):7115.

Beaulieu JM, Leitch IJ, Patel S, Pendharkar A and Knight CA (2008) Genome size is a strong predictor of cell size and stomatal density in angiosperms. New Phytol 179(4):975–986.

Becking T, Gilbert C and Cordaux R (2020) Impact of transposable elements on genome size variation between two closely related crustacean species. Analytical Biochem 600:113770.

Bourque G, Burns KH, Gehring M. et al. (2018) Ten things you should know about transposable elements. Genome Biol 19:199.

Cáceres M, Ranz JM, Barbadilla A, Long M and Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. Science 285(5426):415-418.

Cáceres M, Puig M and Ruiz A (2001) Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. Genome Res 11(8):1353-64.

Casals F, Cáceres M and 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(5):674-685.

Casals F, Cáceres M, Manfrin MH, González J and Ruiz A (2005) Molecular characterization and chromosomal distribution of *Galileo, Kepler* and *Newton*, three *foldback* transposable elements of the *Drosophila buzzatii* species complex. Genetics 169(4):2047-59.

Castle WE, Carpenter FW, Clark AH, Mast SO and Barrows WM (1906) The effects of inbreeding, cross-breeding, and selection upon the fertility and variability of *Drosophila*. Proc Am Acad Arts Sci 41(33):731–786.

Delprat A, Negre B, Puig M and Ruiz A (2009) The transposon *Galileo* generates natural chromosomal inversions in *Drosophila* by ectopic recombination. PLoS One 4(11): e7883.

Dobzhansky T, Ayala FJ, Stebbins GL, Valentine JW (1977) Evolution. WH Freeman, San Francisco, 572 pp.

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

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P and Yang G (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56(9):475-486.

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

Finnegan DJ (2012) Retrotransposons. Curr Biol 22(11):R432-437.

Fontdevila A, Ruiz A, Alonso G and Ocaña J (1981) Evolutionary history of *Drosophila buzzatii*. I. Natural chromosomal polymorphism in colonized populations of the Old World. Evolution 35(1):148–157.

Garcia C, Delprat A, Ruiz A, Valente VL (2015) Reassignment of *Drosophila willistoni* genome scaffolds to chromosome II arms. G3 (Bethesda) 5(12):2559-66.

Gebrie A (2023) Transposable elements as essential elements in the control of gene expression. Mobile DNA 14:9.

Göke J and Ng HH (2016) CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep 17(8):1131-1144.

Gonçalves JW, Valiati VH, Delprat A, Valente VL and Ruiz A (2014) Structural and sequence diversity of the transposon *Galileo* in the *Drosophila willistoni* genome. BMC Genomics 15(1):792.

Goodier JL (2016) Restricting retrotransposons: a review. Mob DNA 7:16.

Gray YH (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet 16(10):461-468

Gregory TR. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76(1):65-101.

Gregory TR (2002) Genome size and developmental complexity. Genetica 115(1):131–146.

Gregory TR (2004) Macroevolution, hierarchy theory, and the C-value enigma. Paleobiology 30(2):179–202.

Gregory TR and Johnston JS (2008) Genome size diversity in the family Drosophilidae. Heredity 101(3):228–238.

Grimaldi DA (1987) Phylogenetics and taxonomy of *Zygothrica* (Diptera: Drosophilidae). Bull Am Mus Nat Hist 186:103–268.

Grimaldi DA (1990) A phylogenetic, revised classification of genera in the Drosophilidae (Diptera). Bull Am Mus Nat Hist 197:1–139.

Guio L and González J (2019). New insights on the evolution of genome content: population dynamics of transposable elements in flies and humans. In: Anisimova M (ed) Evolutionary Genomics. Methods in Molecular Biology, vol 1910. Humana, New York, NY.

Harewood L, Kishore K, Eldridge MD et al. (2017) Hi-C as a tool for precise detection and characterization of chromosomal rearrangements and copy number variation in human tumors. Genome Biol 18:125.

Hayward A and Gilbert C (2022). Transposable elements. Curr Biol 32(17):R904-R909.

Jeffery NW, Ellis EA, Oakley TH and Gregory TR (2017) The genome sizes of ostracod crustaceans correlate with body size and evolutionary history, but not environment. J Hered 108(6):701–706.

Jones RN (2005) McClintock's controlling elements: the full story. Cytogenet Genome Res 109(1-3):90–103.

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.

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115(1):49–63.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ERR, Pelaez J et al. (2021) Highly contiguous assemblies of 101 drosophilid genomes. Elife 10:e66405.

Kobayashi S, Goto-Yamamoto N and Hirochika H (2004) Retrotransposon-induced mutations in grape skin color. Science 304(5673):982.

Lanciano S and Cristofari G (2020) Measuring and interpreting transposable element expression. Nat Rev Genet 21:721–736.

Lim JK and Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. BioEssays 16:269–275.

Liu P, Cuerda-Gil D, Shahid S and Slotkin RK (2022) The epigenetic control of the transposable element life cycle in plant genomes and beyond. Annu Rev Genet 56:63-87.

Mardiros XB, Park R, Clifton B, Grewal G, Khizar AK, Markow TA, Ranz JM and Civetta A (2016) Postmating Reproductive isolation between strains of *Drosophila willistoni*. Fly (Austin) 10(4):162-71.

Markow TA and O'Grady PM (2006) Phylogenetic relationships of Drosophilidae. In: Markow TA and O'Grady PM (eds) *Drosophila*: A Guide to Species Identification and Use. Academic Press, Cambridge, pp. 3–64.

Marzo M, Puig M and Ruiz A (2008) The *foldback*-like element *Galileo* belongs to the *P* superfamily of DNA transposons and is widespread within the *Drosophila* genus. Proc Natl Acad Sci 105(8):2957-62.

Marzo M, Bello X, Puig M, Maside X and Ruiz A (2013) Striking structural dynamism and nucleotide sequence variation of the transposon *Galileo* in the genome of *Drosophila mojavensis*. Mob DNA 4(1):6.

McClintock B (1939) The behavior in successive nuclear divisions of a chromosome broken at meiosis. Proc Natl Acad Sci 25(8):405-416.

McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16:13-47.

McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21(0):197–216.

Mérel V, Boulesteix M, Fablet M and Vieira C (2020) Transposable elements in *Drosophila*. Mob DNA 11:23.

Mills RE, Bennett EA, Iskow RC and Devine SE (2007) Which transposable elements are active in the human genome? Trends Genet 23:183–191.

Morgan TH (1910) Sex limited inheritance in *Drosophila*. Science 32(812):120-122.

Morgan TH (1917) The theory of the gene. Am Nat 51(609):513-544.

Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D and Volff JN (2016) Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect 22(4):312-323.

O'Grady PM and DeSalle R (2018) Phylogeny of the genus *Drosophila*. Genetics 209:1–25.

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

Piégu B, Bire S, Arensburger P and Bigot Y (2015) A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol 86:90-109.

Pita S, Panzera Y, Valente VLS, de Melo ZGS, Garcia C, Garcia ACL, Montes MA and Rohde C (2014) Cytogenetic mapping of the Muller F element genes in *Drosophila willistoni* group. Genetica 142(5):397-403.

Robe LJ, Cordeiro J, Loreto EL and Valente VL (2010) Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila willistoni* subgroup (Diptera: Drosophilidae). Genetica 138(6):601-17.

Rohde C and Valente VL (2012) Three decades of studies on chromosomal polymorphism of *Drosophila willistoni* and description of fifty different rearrangements. Genet Mol Biol 35(4 (suppl)):966-79.

Salzano FM (1956) Chromosomal polymorphism and sexual isolation in sibling species of the *bocainensis* subgroup of *Drosophila*. Evolution 10(3):288-297.

Santos-Colares MC, Valente VLS and Goñi B (2003) The meiotic chromosomes of male *Drosophila willistoni*. Caryologia 56(4):431-437.

Schaack S, Gilbert C and Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25(9):537-546.

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.

Sessegolo C, Burlet N and Haudry A (2016) Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol Lett 12(8):20160407.

Sotero-Caio CG, Platt RN, Suh A and Ray DA (2017) Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol 9:161–177.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell P, Lang M, Courtier-Orgogozo V et al. (2022) Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr Biol 32(1):111-123.

Thomas CA Jr (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237-56.

Throckmorton LH (1975) The phylogeny, ecology, and geography of *Drosophila*. In: King RC (ed) Handbook of Genetics. Plenum Publishing, New York, pp. 421–469.

Valente-Gaiesky VLS. (2019) Can insect assemblages tell us something about the urban environment health? An Acad Bras Cienc 91(supp 3):e20190445.

Wallau GL, Ortiz MF, Loreto EL (2012) Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol Evol 4(8):689-699.

Wallau GL, Capy P, Loreto E, Le Rouzic A and Hua-Van A (2016) VHICA, a new method to discriminate between vertical and horizontal transposon transfer: application to the *Mariner* family within *Drosophila*. Mol Biol Evol 33(4):1094-109.

Waltari E and Edwards SV (2002) Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs. Am Nat 160(5):539–552.

Wang J and Han GZ. A missing link between retrotransposons and retroviruses. mBio 13(2): e0018722.

Wells JN and Feschotte C. (2020) A field guide to eukaryotic transposable elements. Annu Rev Genet 54:539-561.

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci 103(47):17600-17601.

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(12):973-82.

Wicker T, Stritt C, Sotiropoulos AG, Poretti M, Pozniak C, Walkowiak S, Gundlach H and Stein N (2021) Transposable element populations shed light on the evolutionary history of wheat and the complex co-evolution of autonomous and non-autonomous retrotransposons. Adv Genet (Hoboken) 3(1):2100022.

Winge H and Cordeiro A (1963) Experimental hybrids between *Drosophila willistoni* Sturtevant and *Drosophila paulistorum* Dobzhansky and Pavan from southern marginal populations. Heredity 18:215–222.

Winge H (1965) Interspecific hybridization between the six cryptic species of *Drosophila willistoni* group. Heredity 20:9–19.

Yassin, A (2013) Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era. Syst Entomol 38(2), 349–364.

Yusa K, Zhou L, Li MA, Bradley A and Craig NL (2011) A hyperactive *piggyBac* transposase for mammalian applications. Proc Natl Acad Sci 108:1531–1536.

Zanini R, Deprá M and Valente VLS (2015) On the geographic distribution of the *Drosophila willistoni* group (Diptera, Drosophilidae) - updated geographic distribution of the Neotropical *willistoni* subgroup. Dros Inf Serv 98:39-43.

Zanini R, Müller MJ, Vieira GC, Valiati VH, Deprá M and Valente VLS (2018) Combining morphology and molecular data to improve *Drosophila paulistorum* (Diptera, Drosophilidae) taxonomic status. Fly (Austin) 12(2):81-94.

Zeng L, Pederson S, Kortschak R and Adelson DL (2018) Transposable elements and gene expression during the evolution of amniotes. Mobile DNA 9:17.

Zhang SJ, Liu L, Yang R and Wang X (2020) Genome size evolution mediated by *Gypsy* retrotransposons in Brassicaceae. Genomics Proteomics Bioinformatics 18(3):321-332.