UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS GRADUAÇÃO EM GENÉTICA

TESE DE DOUTORADO

Genômica de DNAs repetitivos em *Drosophila virilis*

Autor: Guilherme Borges Dias
Orientador: Dr. Gustavo Campos e Silva Kuhn
Coorientadora: Dr.ª Marta Svartman

Belo Horizonte
2017

043 Dias, Guilherme Borges.
Genômica de DNAs repetitivos em *Drosophila virilis* [manuscrito] / Guilherme Borges Dias. – 2017.

113 f. : il. ; 29,5 cm.

Orientador: Dr. Gustavo Campos e Silva Kuhn. Coorientadora: Dr.ª Marta Svartman.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Genômica - Teses. 2. Mosca-das-frutas. 3. Heterocromatina. 4. Transposons - Teses. 5. Sequências de Repetição em Tandem. I. Kuhn, Gustavo Campos e Silva. II. Svartman, Marta. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575

Ficha elaborada pela Biblioteca do Instituto de Ciências Biológicas da UFMG

Pós-Graduação em Genética
Departamento de Biologia Geral, ICB
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - C.P. 486 - Pampulha - 31270-901 - Belo Horizonte - MG
e-mail: pg-gen@icb.ufmg.br FAX: (+31) - 3409-2570

Pós-Graduação

"Genômica de DNAs repetitivos em Drosophila virilis."

Guilherme Borges Dias

Tese aprovada pela banca examinadora constituída pelos Professores:

Gustavo Campos e Silva Kuhn
UFMG

Antônio Bernardo de Carvalho
UFRJ

Jurandir Vieira de Magalhães
EMBRAPA

Álvaro Gil Araújo Ferreira
Fiocruz

Leonardo Koerich
UFMG

Marta Svartman
UFMG

Belo Horizonte, 22 de setembro de 2017.

Guilherme Borges Dias

Genômica de DNAs repetitivos em *Drosophila virilis*

Tese apresentada ao Programa de Pós-Graduação em Genética do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética.

Orientador: Dr. Gustavo Campos e Silva Kuhn
Coorientadora: Dr.ª Marta Svartman

Belo Horizonte
2017

Agradecimentos

Aos meus pais Valdete e David, pelo apoio, incentivo, e amor incondicional.

Ao Professor Gustavo Kuhn, pela dedicação ao longo dos anos e liberdade na escolha do projeto.

À Prof.ª Marta Svartman por todos os ensinamentos, conselhos e amizade.

Aos amigos que fazem as dificuldades parecerem menos ameaçadoras e o dia a dia mais feliz.

A todos com os quais convivi neste trajeto e me ajudaram a aprender lições importantes e crescer acadêmica e pessoalmente.

Sumário

Lista de Figuras

Capítulo 1

Capítulo 2

Capítulo 3

Lista de Tabelas

Lista de abreviaturas

AIC	Akaike Information Criterion
bp	base pairs
CTR	Central Tandem Repeats
DINEs	*Drosophila* INterspersed Elements
FISH	Fluorescence *in situ* hybridization
Gb	Gigabases
IR	Inverted Repeat
kb	kilobases
Mbp	Megabases
miRNA	micro RNA
MSL	Male-specific lethal complex
pb	pares de base
pg	picograma
piRNA	piwi-interacting RNA
satDNA	satellite DNA
siRNA	small interfering RNA
SPR	Subtree pruning and regrafting
subTIRs	Subterminal Inverted Repeats
TE	Transposable Element
TR	Tandem Repeats

Resumo

Os DNAs repetitivos representam uma grande fração dos genomas eucariotos. Estes elementos são sub-representados em análises genômicas devido às dificuldades computacionais de lidar com repetições, especialmente pelo tamanho curto das *reads* de sequenciamento comumente utilizadas. *Drosophila virilis* é a espécie de *Drosophila* com o maior genoma e maior proporção de DNAs repetitivos entre as espécies sequenciadas do gênero. Esta tese reúne esforços de caracterização de DNAs repetitivos em *D. virilis*. No primeiro capítulo, caracterizamos uma repetição em tandem descrita como a mais abundante em *D. virilis* e discutimos sua origem, distribuição cromossômica e impacto na evolução do genoma. Demonstramos que se trata de um DNA satélite que surgiu a partir da amplificação de repetições internas de um transposon do tipo *Helitron*. O mesmo processo ocorreu independentemente a partir de um *Helitron* homólogo em *D. biarmipes*, uma espécie filogeneticamente distante de *D. virilis*. Em *D. virilis*, este transposon e as repetições a ele associadas são abundantes em regiões de transição de cromatina e produzem piRNAs, indicando possíveis efeitos destas sequências na modulação da cromatina. No segundo capítulo discutimos o possível papel destas repetições derivadas de transposons na modulação da cromatina em todo o genoma. Também descrevemos modelos de transposição que abordam a frequente inserção de *Helitrons* em tandem, um fenômeno até então não discutido na literatura. No terceiro e último capítulo caracterizamos uma repetição em tandem de 172 pb em *D. virilis*. Descobrimos que esta repetição é um minissatélite extremamente abundante (em torno de 15,500 cópias) e presente em todas as espécies do subgrupo. Também identificamos a presença de pequenos RNAs derivados deste minissatélite em embriões e gônadas de *D. virilis*. Utilizando *reads* longas de sequenciamento de terceira geração, construímos uma nova montagem genômica para *D. virilis*. Esta montagem se mostrou muito mais íntegra do que o genoma de referência (sequenciado pelo método de Sanger), e com ela foi possível recuperar mais cadeias completas contendo esta repetição. Os resultados obtidos nesta tese representam uma importante adição à caracterização de DNAs repetitivos no genoma de *D. virilis*, uma espécie tradicionalmente utilizada em estudos comparativos no gênero *Drosophila*.

Abstract

Repetitive DNAs represent a large fraction of eukaryote genomes. These elements are underrepresented in genomic analyses because of the computational limitations in dealing with repeats, especially with current short read sequencing technologies. *Drosophila virilis* is the *Drosophila* species with the largest genome and largest proportion of repetitive DNA among all the sequenced species of the genus. This thesis reunites efforts in characterizing repetitive DNAs of *D. virilis*. In the first chapter, we characterized a tandem repeat identified as the most abundant in *D. virilis* and discussed its origin, chromosome distribution and impact in genome evolution. We demonstrate that this repeat is a satellite DNA that emerged from internal tandem repeats from an *Helitron* transposon. The same phenomenon happened independently from a homologous *Helitron* in *D. biarmipes*, a phylogenetically distant species. In *D. virilis*, this transposon and its associated repeats are abundant in chromatin transition zones and generate piRNAs, suggesting possible chromatin modulation roles for these sequences. In the second chapter, we discuss the possible role of transposon derived repeats on genome-wide chromatin modulation. We also describe transposition models which explain the frequent tandem insertions of *Helitrons*, a phenomenon that has not been addressed so far. In the third and last chapter, we characterized a 172 bp tandem repeat in *D. virilis*. We determined that this repeat is an extremely abundant minisatellite (around 15500 copies) that is present in all species from the virilis subgroup. We also identified small RNAs derived from these repeats in embryos and gonads of *D. virilis*. By using long reads from third generation sequencing we produced a new genome assembly for *D. virilis*. This assembly proved to be a lot more contiguous than the reference genome (sequenced with the Sanger method) and recovered more complete arrays of the 172 bp tandem repeats. The results presented herein represent important additions to the characterization of repetitive DNA in the genome of *D. virilis*, a species traditionally used in comparative studies of the *Drosophila* genus.

1. Introdução

A quantidade de DNA no genoma haploide de um organismo, ou valor C, varia enormemente entre os seres vivos e não está correlacionada à complexidade biológica. Organismos "simples" frequentemente contêm mais DNA do que organismos "complexos" e mesmo espécies próximas podem apresentar grande variação no valor C. De modo geral, os organismos possuem muito mais DNA do que o esperado de acordo com seu número de genes e complexidade do desenvolvimento (Gregory 2001). Por exemplo, o gafanhoto *Podisma pedestris* possui um valor C aproximadamente sete vezes maior que o do morcego *Eptesicus brasiliensis*, com 16 e 2.3 gigabases (Gb), respectivamente (Westerman et al. 1987; Smith et al. 2013). O eucarioto unicelular *Paramecium tetraurelia* possui um valor C de ~0,1 pg e tem 40.000 genes, enquanto vertebrados como o cão possuem valor C de ~3,2 pg e possuem em torno de 20.000 genes (Aury et al. 2006; Lindblad-Toh et al. 2005). A princípio, estas discrepâncias foram vistas como paradoxais. O paradoxo do valor C foi parcialmente solucionado quando se notou que grande parte dos genomas eucariotos correspondia a DNA repetitivo não codificante de proteínas e que a maior parte da variação no valor C poderia ser explicada por diferentes quantidades deste DNA repetitivo. De modo geral, genomas maiores abrigam uma maior quantidade de repetições do que genomas menores (Gregory 2001). O genoma humano, por exemplo, contém 69% de DNAs repetitivos (de Koning et al. 2011). Em comparação, os genes codificadores de proteínas, de RNAs, e as sequências reguladoras perfazem apenas 5,1% do genoma humano (Strachan & Read 2010).

Apesar da grande contribuição dos DNAs repetitivos no tamanho dos genomas, as análises funcionais e evolutivas tradicionalmente têm se focado nos genes codificadores de proteínas, uma vez que estes elementos possuem papéis biológicos mais óbvios, apreciados há mais tempo, e existem mais ferramentas disponíveis para sua análise. Apesar de pouco explorados se comparados à fração gênica, os DNAs repetitivos têm ganhado atenção em estudos de organização e função do genoma (Kuhn 2015; Maumus et al. 2015).

1.1. Principais tipos de DNA repetitivo

Existem várias classes de DNAs repetitivos nos genomas eucariotos. De forma geral, eles podem ser divididos em repetições dispersas e repetições em tandem (Figura 1). O tipo mais abundante de repetições dispersas são os elementos transponíveis (também chamados de transposons ou TEs, sigla em inglês para *Transposable Elements*). As repetições em tandem incluem micro e minissatélites e os DNA satélites (ou *satellite DNAs*, satDNAs).

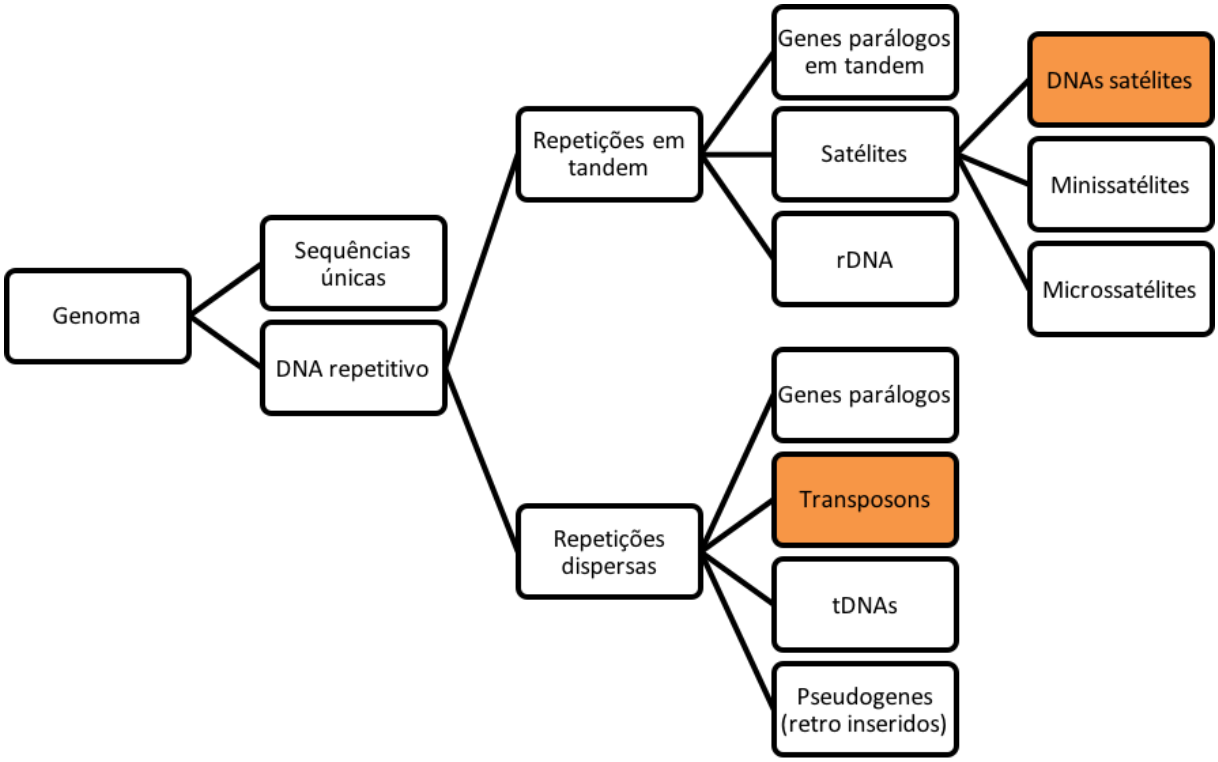


Figura 1. Esquema simplificado da organização dos genomas eucariotos, com destaque para os tipos de DNAs repetitivos mais abundantes. rDNA, genes que codificam os RNAs ribossômicos; tDNAs, genes que codificam os RNAs de transferência. (Adaptado de Richard et al. 2008).

Os TEs são elementos genéticos com capacidade intrínseca de propagação no genoma. Eles podem ser autônomos, caso codifiquem as enzimas responsáveis por sua mobilização, ou não-autônomos, quando utilizam as enzimas codificadas por outros TEs. Os genomas eucariotos abrigam uma enorme quantidade e diversidade de TEs, que apenas recentemente começou a ser explorada em profundidade. Podemos dividir estes elementos em dois grupos principais: os retroelementos, que são TEs que utilizam uma sequência intermediária de RNA na transposição; e os transposons de DNA, que não utilizam intermediário de RNA (Finnegan 1989).

Dentro da classe de transposons de DNA existe um grupo de elementos recentemente descobertos e pouco conhecidos, os Helitrons (Kapitonov and Jurka 2001). Estes elementos não possuem uma estrutura simétrica e sua transposição parece ser mediada por uma transposase do tipo Rep-Helicase, similar à replicação de plasmídeos e elementos de inserção bacterianos (Kapitonov and Jurka 2007).

Além de se mobilizarem dentro do genoma, TEs de uma espécie podem invadir espécies diferentes. O fenômeno de invasão de um novo genoma por um TE é chamado de HTT (*Horizontal Transposon Transfer*) e foi relatado em diversos grupos de eucariotos

(Dotto et al. 2015). Este tipo de fenômeno contribui para o surgimento de novidades evolutivas e parece fazer parte do ciclo de vida dos TEs, assegurando sua perpetuação (Hua-Van et al. 2011).

Dentre as repetições em tandem se destacam os micro e minissatélites, e os DNAs satélites (Richard et al. 2008). Microssatélites são repetições em tandem com monômeros de 1-10 pares de base (pb) e cadeias geralmente com no máximo algumas centenas de pb. Já os minissatélites normalmente possuem monômeros com até 100 pb e cadeias de 0.5 a 30 kilobases (kb; Armour and Jeffreys 1992; Charlesworth et al. 1994). Micro e minissatélites ocorrem dispersos ao longo dos cromossomos, com possível acúmulo dos minissatélites nas regiões distais (Tautz 1993). Uma vez que os micro e minissatélites são principalmente usados na identificação de indivíduos e como marcadores populacionais, estas sequências são quase sempre discutidas no contexto de *loci* individuais, e análises genômicas e evolutivas envolvendo essas repetições são escassas. Os DNAs satélites podem ter monômeros com desde poucos pb até alguns kb e normalmente ocorrem em longas cadeias que correspondem aos blocos de heterocromatina (Plohl et al. 2008). Entretanto, cadeias curtas podem existir dispersas na eucromatina (Kuhn et al. 2012; Brajkovic et al. 2013).

As várias cópias de uma mesma família de repetições em tandem sofrem homogeneização gerando um padrão de identidade de sequência espécie-específico chamado de evolução combinada (*Concerted Evolution*; Liao 1999). Este padrão é gerado por vários mecanismos de trocas não-recíprocas chamados conjuntamente de 'impulso molecular' (*molecular drive*; Dover 1982). Estes mecanismos também geram variação no número de cópias dos DNA satélites e incluem o *crossing-over* desigual, conversão gênica, derrapagem da polimerase, transposição e amplificação de DNA extracromossômico por círculo rolante (Hourcade et al. 1973; Dover 1982, 2002; Charlesworth et al. 1994).

1.2. Relações evolutivas entre TEs e DNAs satélites

Apesar de terem organização e distribuição genômica predominantemente distintas, TEs e DNA satélites podem ter relações de parentesco evolutivo. A simples similaridade de sequências entre TEs e alguns DNA satélites é indicativo dessa relação, mas apenas uma análise detalhada pode fornecer informações sobre os mecanismos e consequências desta dinâmica evolutiva e a direção do processo. Casos em que TEs supostamente contribuíram para o surgimento de novos DNA satélites são frequentemente relatados (Kapitonov et al. 1998; Macas et al. 2009; Satovic & Plohl 2013; Sharma et al. 2013). O processo inverso, com DNAs satélites contribuindo na formação de novos TEs é mais raro (Heikkinen et al. 1995).

Ainda há poucas discussões sobre aspectos gerais destes fenômenos, como mecanismos ou regiões cromossômicas preferenciais e o impacto destes processos para a evolução de genomas eucariotos. Recentemente, Dias et al. (2014) sugeriram que as regiões de transição entre heterocromatina e eucromatina, chamadas de β-heterocromatina, poderiam ser *hotspots* para o surgimento de novos DNA satélites a partir de TEs residentes (Figura 2). Entretanto, estas regiões de transição são mais facilmente identificadas nos cromossomos politênicos de *Drosophila*, de modo que dados de espécies que não possuem cromossomos politênicos não podem ser facilmente comparados.

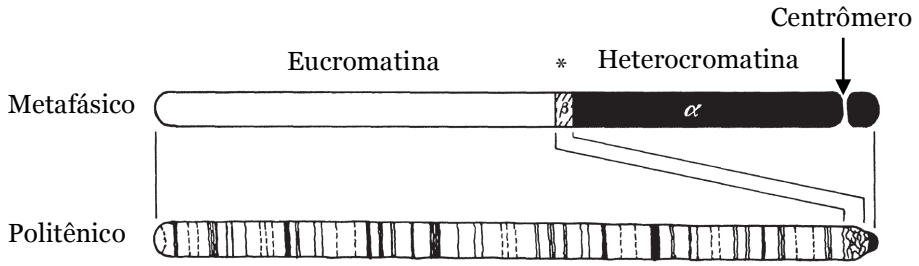


Figura 2. Comparação de um cromossomo metafásico e seu correspondente após politenização. No cromossomo politênico, a heterocromatina centromérica e pericentromérica permanece não replicada. Já a zona de transição (β-heterocromatina, marcada com um asterisco) sofre replicação se apresentando como uma massa de DNA sem o padrão de bandas característico da eucromatina politênica. Fonte: Gall 1973.

1.3. Papéis biológicos de DNAs repetitivos

Apesar de tradicionalmente considerados 'DNA entulho', várias descobertas mostraram que satélites e TEs podem ter papéis biológicos importantes. A alta frequência de longas cadeias de DNA satélites nos centrômeros eucariotos sugere um papel para estas sequências na função centromérica (Masumoto et al. 2006). Além disso, centrômeros compostos de TEs e/ou uma mistura de DNA satélites e TEs também foram encontrados (Plohl et al. 2014). Apesar disso, há relatos de neocentrômeros funcionais isentos de qualquer DNA repetitivo (Gong et al. 2012). Uma hipótese é a de que o acúmulo de DNA satélites e/ou TEs faz parte do processo de amadurecimento do centrômero, conferindo mais estabilidade e assegurando a divisão correta dos cromossomos (Piras et al. 2010).

Além da manutenção centromérica, TEs e satélites podem estar associados à modulação da cromatina e regulação gênica (Ugarkovic 2005; de Souza et al. 2013). Por exemplo, o TE *Hopscotch* se inseriu na região reguladora do gene *teosinte branched1* do milho (*Zea mays*) e passou a atuar como um *enhancer,* gerando a superexpressão desse gene, o que foi um dos passos chave na domesticação desta espécie a partir de seu ancestral

(Studer et al. 2011). Os micro e minissatélites também são importantes componentes da regulação gênica em eucariotos por sua frequente variação em número de cópias. É estimado que até 20% de todos os genes e promotores em eucariotos possuam trechos instáveis dessas repetições e, além de causarem doenças como a síndrome do cromossomo X frágil, essa variação também resulta em plasticidade fenotípica (Gemayel et al. 2010).

Inserções individuais de TEs também podem influenciar a expressão gênica local por meio de uma alteração do estado da cromatina. Isto se dá pela interação destes TEs com RNAs curtos complementares que, por sua vez, recrutam moduladores de cromatina e geram heterocromatinização local (Sentmanat & Elgin 2012; Lee 2015). Um efeito similar de heterocromatinização é observado para cadeias de repetições em tandem e, apesar de ainda ser pouco compreendido, este efeito também foi atribuído à interação com RNAs de interferência (Martiessen 2003).

1.4. DNAs repetitivos em *Drosophila*: o modelo *Drosophila virilis*

Moscas do gênero *Drosophila* têm sido utilizadas como modelos em laboratório por mais de 100 anos e proporcionaram importantes avanços em diversas áreas da biologia (Markow 2015). Em 2007, um grande passo foi dado com a publicação de dados genômicos para 12 espécies de *Drosophila* (*Drosophila* 12 Genomes Consortium), colocando *Drosophila* na vanguarda da genômica comparativa. Dentre estas 12 espécies, há representantes dos principais grupos filogenéticos de *Drosophila*, compondo uma importante amostra da diversidade genética e ecológica do gênero (Figura 3; Crosby et al. 2007).

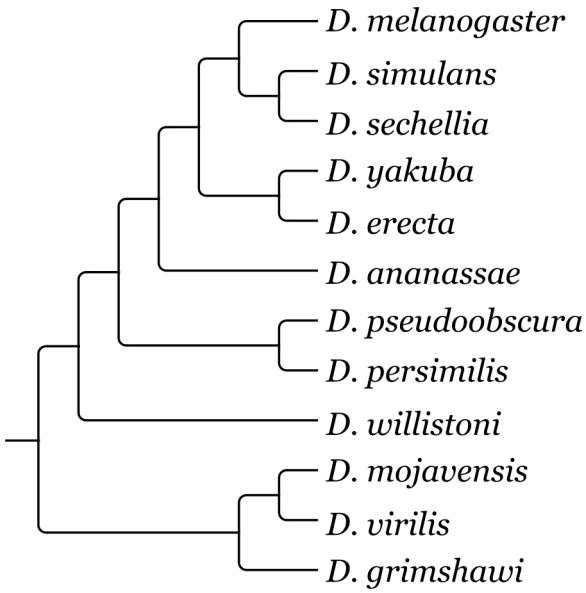


Figura 3. Filogenia das 12 espécies de *Drosophila* com genoma sequenciado reportadas em 2007 (*Drosophila* 12 Genomes Consortium).

Das 12 espécies inicialmente sequenciadas, *D. virilis* é a que possui o maior genoma e a maior proporção de DNA satélites (Figura 4; Bosco et al. 2007). Com um genoma de aproximadamente 360 Mb, perfaz o dobro do genoma de *D. melanogaster* (aprox. 180 Mb). A proporção de heterocromatina também é maior em *D. virilis* (50%; Mahan & Beck 1986) do que em *D. melanogaster* (30%; Gatti & Pimpinelli 1992). Em alguns estudos, o tamanho do genoma vem sendo correlacionado com características biológicas, como tamanho corporal e taxa de desenvolvimento (Gregory 2005). De fato, *D. virilis* é maior e apresenta um período de desenvolvimento de ovo a adulto mais longo do que a maior parte das espécies de *Drosophila* estudadas (Gregory & Johnston 2008). Estas características, juntamente com a sua alta resistência e fácil manutenção em laboratório, tornam *D. virilis* um modelo interessante para estudos sobre DNAs repetitivos, evolução do genoma e flutuações do valor C.

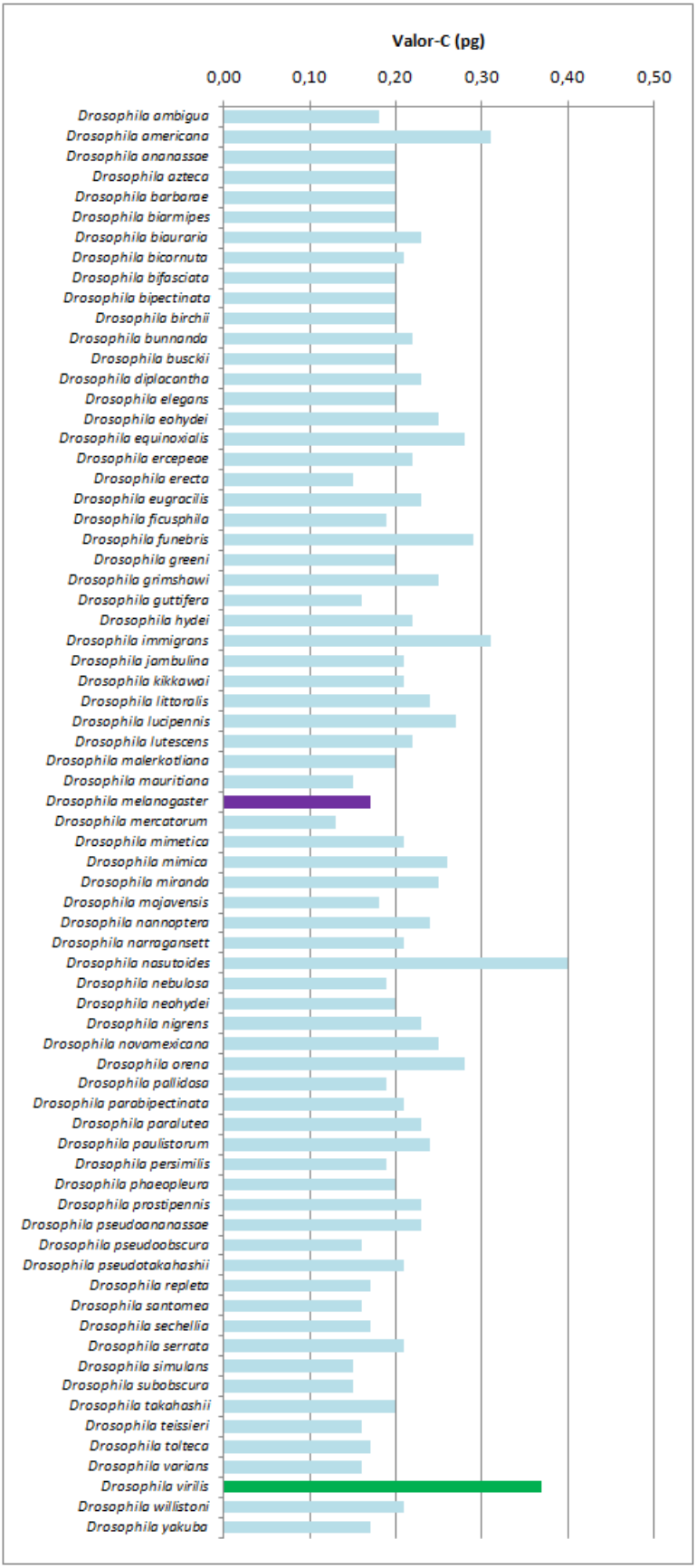


Figura 4. Estimativas de tamanho do genoma em picogramas (pg) para espécies do gênero *Drosophila* com destaque para *D. melanogaster* e *D. virilis* (Dados disponíveis em http://www.genomesize.com). 1 pg equivale a aproximadamente 978 Mb (Dolezel et al. 2003).

Mais de 800 famílias de TEs foram identificadas no genoma montado de *D. virilis* (Feschotte et al. 2009). Entretanto, dados detalhados sobre estrutura, distribuição cromossômica e dinâmica evolutiva destes elementos não foram investigados. Em contraste, apenas cinco DNA satélites foram descritos em *D. virilis*. Ainda nos anos 1970, foram identificados três DNA satélites heptanucleotídeos chamados de satélites I, II e III (Gall et al. 1971). Estas sequências foram identificadas através de centrifugação em um gradiente de densidade com cloreto de césio e estimou-se que sua abundância representava ~40% do genoma de *D. virilis*. A hibridização *in situ* dos DNA satélites I e II evidenciou acúmulo destas sequências nos blocos heterocromáticos de todos os cromossomos, com exceção do Y (Gall et al. 1971).

O quarto satDNA identificado, denominado pvB370, possui monômeros de ~370 pb similares às extremidades dos transposons pDv. A distribuição filogenética do satDNA pvB370 e dos TEs pDv indica que o satDNA é mais antigo e, portanto, pode ter participado da formação dos TEs (Heikkinen et al. 1995). Recentemente, um quinto satDNA foi identificado em *D. virilis*. A análise destas sequências indicou que este satDNA surgiu da expansão de repetições em tandem pré-existentes de um TE do tipo *foldback*, denominado de *Tetris* (Dias et al. 2014).

Apesar da caracterização dos DNA satélites de *D. virilis* já ter sido parcialmente feita, o genoma montado desta espécie representa apenas metade do tamanho estimado por citometria de fluxo, e inclui principalmente a fração eucromática e com menor densidade de repetições (*Drosophila* 12 Genomes Consortium, 2007). Isto indica que há ainda muitas famílias de DNAs repetitivos a serem descobertas em *D. virilis*. De fato, numa análise recente, identificou-se o que seria a repetição em tandem mais abundante em mais de 200 genomas eucariotos (Melters et al. 2013), incluindo *D. virilis*. Esta análise identificou em *D. virilis* um sexto satDNA de aproximadamente 150 pb, que não foi caracterizado experimentalmente. Além disso, uma nova família de repetições em tandem foi identificada por Abdurashitov et al. (2013). Estas repetições possuem um consenso de 172 pb e aparentemente estão dispersas no genoma, mas nenhuma caracterização adicional foi feita.

2. Justificativa

Os DNAs repetitivos frequentemente representam a maior parte do genoma de eucariotos. Apesar disto, sua identificação e caracterização ocorrem num ritmo muito mais lento do que o da fração gênica do genoma. Ainda que boa parte dos DNAs repetitivos possa ser material "inerte" nos genomas em que residem, já há vários casos descritos em que elementos repetitivos assumiram papeis funcionais (Biscotti et al. 2015; Jangam et al. 2017).

Parte da dificuldade de análise dos DNAs repetitivos deriva da qualidade atual das montagens genômicas, que são altamente fragmentadas e frequentemente excluem boa parte das regiões repetitivas (Treangen & Salzberg 2012). Desta forma, a utilização complementar de métodos *in silico* que não dependem de montagem e a integração de métodos experimentais de mapeamento cromossômico são indispensáveis na caracterização destes elementos.

A espécie *D. virilis* se destaca dentro do gênero *Drosophila* por conter um dos maiores genomas e uma das maiores proporções de DNAs repetitivos. Apesar de ter tido seu genoma publicado em 2007, poucas famílias de DNAs repetitivos foram caracterizadas em *D. virilis* desde então.

3. Objetivos

Esta tese sumariza esforços de caracterização dos DNAs repetitivos em *D. virilis*.

Objetivos por capítulo:

Capítulo 1

- Caracterizar a repetição em tandem descrita como a mais abundante no genoma de *D. virilis* por Melters et al. (2013).

Capítulo 2

- Discutir o impacto da expansão de repetições em tandem a partir de TEs sobre a modulação da cromatina no genoma.
- Propor modelos de transposição que abordem o recorrente fenômeno de inserções em tandem de *Helitrons*.

Capítulo 3

- Gerar uma nova montagem genômica para *D. virilis* utilizando dados de sequenciamento de terceira geração.
- Utilizar os recursos genômicos disponíveis e a nova montagem para caracterizar uma família de repetições em tandem identificadas por Abdurashitov et al. (2013).

Capítulo 1

Helitrons shaping the genomic architecture of *Drosophila*: enrichment of *DINE-TR1* in α-and β-heterochromatin, satellite DNA emergence, and piRNA expression

Este capítulo é composto de um artigo publicado na revista *Chromosome Research*. O artigo descreve a identificação de um subgrupo de TEs pertencente ao grupo dos *Helitrons* e chamado de *DINE-TR1*. Estes transposons possuem repetições em tandem internas de 150 pb, parcialmente conservadas mesmo entre espécies de Diptera que divergiram do ancestral comum há mais de 60 milhões de anos atrás. O objetivo inicial deste trabalho era o de caracterizar uma repetição em tandem de 150 pb que foi identificada como a mais abundante em *Drosophila virilis*. Descobrimos que essa repetição era na verdade derivada do TE *DINE-TR1* e que sofreu amplificações independentemente em pelo menos duas espécies de *Drosophila*, *D. virilis* e *D. biarmipes*, resultando no que podem ser considerados novos DNAs satélites. Também detectamos a presença de *DINE-TR1* em regiões cromossômicas produtoras de piRNA, bem como a presença de RNAs curtos derivados de *DINE-TR1* nos tecidos gonadais e embrionários de *D. virilis*. Discutimos o potencial impacto de *DINE-TR1* no genoma de *D. virilis* e de outros Diptera.

Helitrons shaping the genomic architecture of *Drosophila*: enrichment of *DINE-TR1* in α and β-heterochromatin, satellite DNA emergence and piRNA expression

Guilherme B. Dias, Pedro Heringer, Marta Svartman, Gustavo C.S. Kuhn*

Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

Running title: *DINE-TR1 Helitrons* in *Drosophila*

*Corresponding author: Gustavo Kuhn, Departamento de Biologia Geral Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, telephone number: +55 (31) 3409-2599, fax number: +55 (31) 3409-2567, e-mail address: gcskuhn@ufmg.br

Abstract

DINEs (*Drosophila* INterspersed Elements) constitute an abundant but poorly understood group of *Helitrons* present in several *Drosophila* species. The general structure of *DINEs* includes two conserved blocks that may or not contain a region with tandem repeats in between. These central tandem repeats (CTRs) are similar within species but highly divergent between species. It has been assumed that CTRs have independent origins. Herein we identify a subset of *DINEs*, termed *DINE-TR1*, which contain homologous CTRs of approximately 150 bp. We found *DINE-TR1* in the sequenced genomes of several *Drosophila* species and in *Bactrocera tryoni* (Acalyptratae, Diptera). However, interspecific high sequence identity (~88%) is limited to the first ~30 bp of each tandem repeat, implying that evolutionary constraints operate differently over the monomer length. *DINE-TR1* is unevenly distributed across the *Drosophila* phylogeny. Nevertheless, sequence analysis suggests vertical transmission. We found that CTRs within *DINE-TR1* have independently expanded into satellite DNA-like arrays at least twice within *Drosophila*. By analyzing the genome of *D. virilis* and *D. americana*, we show that *DINE-TR1* is highly abundant in pericentromeric heterochromatin boundaries and some telomeric regions and in the Y chromosome. It is also present in the centromeric region of one autosome from *D. virilis* and dispersed throughout several euchromatic sites in both species. We further found that *DINE-TR1* is abundant at piRNA clusters and small *DINE-TR1*-derived RNA transcripts (~25 nt) are predominantly expressed in testes and ovaries, suggesting active targeting by the piRNA machinery. These features suggest potential piRNA-mediated regulatory roles for *DINEs* at local and genome-wide scales in *Drosophila*.

Key words: *Drosophila virilis*, *Drosophila americana*, transposable element, genome evolution, *Helitrons*

Introduction

Satellite DNAs (satDNA) and transposable elements (TEs) are the main components of heterochromatin and important contributors to the genome architecture and evolution of eukaryotes (Heslop-Harrison and Schwarzacher 2011; Wallrath et al. 2014). These repetitive sequences participate in several biological processes, such as the formation and propagation of heterochromatin (Volpe et al. 2002; de Wit et al. 2005; Vermaak and Malik 2009), centromere maintenance and function (Malik and Henikoff 2009; Ugarkovic 2009; Vermaak and Malik 2009; Plohl et al. 2014; Rošić et al. 2014) and gene expression regulation (Volpe et al. 2002; Menon et al. 2014). They also contribute to the evolution of adaptive traits and the establishment of reproductive barriers between species (Gregory and Johnston 2008; Ferree and Barbash 2009; Brown and O'Neill 2010; Feliciello et al. 2015). There is growing evidence revealing evolutionary connections between TEs and satDNAs in many organisms. In this context, TEs appear to be an important source for satDNA origin, either by creating tandem repeats by ectopic recombination or by the amplification of pre-existing internal repeat motifs (e.g. Macas et al. 2009; Brajkovic et al. 2013; Satovic and Plohl 2013; Dias et al. 2014).

Regardless of their functional roles, the indiscriminate spread of repetitive sequences is usually selected against because of the potential deleterious consequences caused by ectopic recombination. The strength of the purifying selection varies according to recombination rates, copy numbers and elements lengths (Petrov et al. 2011). At the molecular level, one of the main mechanisms of defense against transposition events in germ line cells is the expression of piRNAs, which interact with proteins from the Piwi clade to specifically target and silence TEs and other repeats (Kalmykova et al. 2005; Saito et al. 2006; Brennecke et al. 2007; Aravin et al. 2007; Grimson et al. 2008). The majority of piRNAs in *Drosophila melanogaster* derives from variable sized clusters that lack protein-coding genes and are replete with eroded remnants of ancient TE insertions and other repeats (Aravin et al. 2007; Brennecke et al. 2007; Jordan and Miller 2008).

Several examples show that TE and satDNA abundance relates with genome size in a wide range of organisms (Kidwell 2002; Gregory and Johnston 2008; Bosco et al. 2007). In this respect, *Drosophila virilis* is of particular interest because it has an atypical large genome with 360-440 Mb (almost twice as big as the genome of many other *Drosophila* species, including *D. melanogaster*) and one of the highest contents of heterochromatin within the genus (50% in contrast to ~30% in *D. melanogaster*) (Gatti et al. 1976; Bosco et al. 2007). Moreover, the availability of the *D. virilis* sequenced genome (*Drosophila* 12 Genomes Consortium 2007) provides a great opportunity to study the impact of repetitive DNAs in genome architecture and evolution.

TEs account for ~14% of the *D. virilis* genome (*Drosophila* 12 Genomes Consortium 2007) and *DINE*s (*Drosophila* INterspersed Elements; Locke et al. 1999), with over 3,000 copies, are among the most abundant ones (Yang and Barbash 2008). *DINEs* are elusive Dipteran transposons from the *Helitron* group thought to mobilize via a rolling circle mechanism (Kapitonov and Jurka 2001; Kapitonov and Jurka 2007b; Yang and Barbash 2008). Their general structure includes subterminal inverted repeats (subTIRs), a short inverted repeat (IR), a core region conserved between elements from distantly related species, a microsatellite region of variable size, a central region with tandem repeats (CTRs) and a stem loop at the 3' end (Fig. 1A; Yang and Barbash 2008; Thomas et al. 2014).

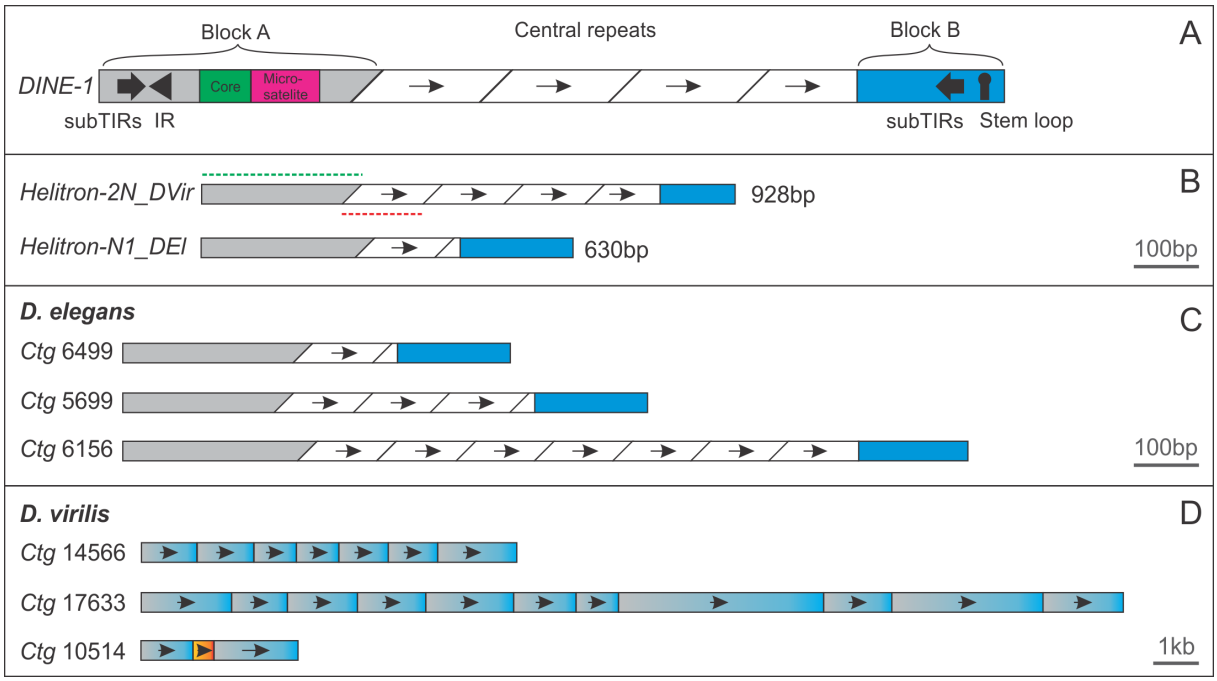


Fig. 1. (*A*) General organization of *DINE-1* elements. Redrawn from Thomas et al. (2014). (*B*) *Helitron* consensus identified as *DINE-TR1* from *D. virilis* and *D. elegans*. Dashed lines above block A and bellow the central repeats indicate segments used as probes in the FISH experiments. (C) Representation of typical *DINE-TR1* elements found in *D. elegans* contigs. (D) Representatives of *D. virilis DINE-TR1* tandem insertions including a *DINE* from a distinct group (*Helitron-1*) in between two *DINE-TR1*s (*Helitron-2*).

*DINE*s are abundant at heterochromatic and euchromatic regions, including several insertions in or around genes (Yang and Barbash 2008). In *D. melanogaster* and *D. simulans*, *DINE*s are frequently found inserted near cytochrome P450 genes related to insecticide resistance, which could indicate a possible regulatory role (Carareto et al. 2014). In *D. miranda*, dosage compensation of the neo-X chromosome was achieved through the

co-optation of *DINE*-related *Helitrons* that recruit the male-specific lethal (MSL) complex (Ellison and Bachtrog 2013). A large survey of *DINE* elements in 12 sequenced *Drosophila* genomes revealed that the evolution of these elements is highly dynamic and include recent transpositional bursts in several species. The *DINE*-CTRs were also found to be similar within species but very divergent between species and it has been assumed that they have independent origins in the 12 analyzed *Drosophila* species (Yang and Barbash 2008). Despite their perceivable importance in shaping the *Drosophila* genome, our knowledge about *DINE*s is still very limited.

SatDNAs account for ~45% of the *D. virilis* genome (Bosco et al. 2007) and most of the heterochromatin present in this species consists of three abundant homologous satDNAs (satellites I, II and III) displaying heptanucleotide repeat units. Despite the fact that these three satellites account for ~40% of the *D. virilis* genome (Gall et al. 1971), a recent bioinformatic survey in several eukaryotic sequenced genomes identified a 150 bp sequence as the most abundant tandem repeat (TR) of *D. virilis*, potentially corresponding to the DNA underlying the centromeres (Melters et al. 2013). Abdurashitov et al. (2013), using in *silico* and *in vitro* DNA digestion, independently identified the same 150 bp repeat as part of a *Helitron* called *Helitron-2_DVir*.

In the present work, we aimed to investigate the association between the 150 bp TR and *Helitron-2_DVir* and to determine their distribution, organization and impact in the *Drosophila* genome. We found that the *Helitron-2_DVir* containing 150 bp repeats is part of a subgroup of *DINE*s that we called *DINE-TR1*. In contrast to what was previously assumed for *DINEs*, the CTRs from *DINE-TR1* share homology among several species. Our study revealed that *DINE-TR1* is restricted to Acalyptratae (Diptera), but display a patchy distribution within the *Drosophila* genus. After analyzing the chromosomes of *D. virilis* and its closely related species *D. americana*, we found that *DINE-TR1* is highly abundant at several genomic regions. Analysis of the *D. virilis* small RNA profile pointed to the involvement of *DINE-TR1* in piRNA expression. We discuss how our findings shed light on the role played by *DINE*s in several aspects of *Drosophila* genome architecture and evolution.

Material and methods

Identification and phylogenetic distribution of *DINE-TR1*

DINE-TR1 was initially identified in *D. virilis* and *D. elegans* after sequence comparisons between the most abundant TRs identified in 21 *Drosophila* species with sequenced genomes reported by Melters et al. (2013). In order to identify *DINE-TR1* related elements in other *Drosophila* species we performed a series of searches. First, we used the Tandem Repeats Finder software (Benson 1999) to look for CTRs in all *Helitrons* from *Drosophila* available in Repbase (Jurka et al. 2005). In addition, dot plots were used for visualization of the organization and size of the arrays (Junier and Pagni 2000). Next, we compared these elements with *Helitron-2_DVir* (a *DINE-TR1* from *D. virilis*) through dot plots in order to identify similarity between the CTRs from *DINE-TR1*.

The remaining *Drosophila* species with available sequenced genomes but without any *Helitron* consensus available at Repbase were queried using the *DINE-TR1* consensus from the available closest species. Besides the sequenced genomes available at Flybase (http://flybase.org), we searched for *DINE-TR1* in other species with recently sequenced genomes, including *D. americana* (http://cracs.fc.up.pt/~nf/dame/), *D. suzukii* (available in NCBI) and *D. buzzatii* (http://dbuz.uab.cat/) (Fonseca et al. 2013; Ometto et al. 2013; Guillén et al. 2014).

To assess the phylogenetic distribution of *DINE-TR1* we also performed searches in the sequenced genomes of other Diptera, including *Bactrocera tryoni*, *Lucilia cuprina*, *Musca domestica* and *Glossina morsitans*, all available at NCBI (Gilchrist et al. 2014; Scott et al. 2014; International *Glossina* Genome Initiative 2014).

Multiple Sequence Alignments (MSAs) and phylogenetic reconstruction

MSAs were performed using the M-Coffee web-server with the default options (Tommaso et al. 2011), and visualized and edited in Jalview (Waterhouse et al. 2009). Maximum Likelihood phylogenies were estimated using PhyML (Guindon and Gascuel 2003) with the best substitution model and parameters according to the Akaike Information Criterion (AIC) as determined by JModeltest version 2.1.4 (Darriba et al. 2012). Trees were reconstructed using the Subtree Prunning and Regrafting algorithm (SPR) and statistical support was calculated after 1000 bootstrap replicates.

Fluorescence *in situ* Hybridizations

Mitotic metaphases were obtained from the neuroblasts of wandering third instar larvae from *D. virilis* (strain 15010-1051.51) and *D. americana* (strain W11) according to the method described in Baimai (1977). Polytene chromosomes were prepared following the acetic acid squash protocol (Ashburner 1989). Specific probes were obtained from *D. virilis* by PCR with primers for *DINE-TR1* block A (forward 5' TTATACCCTTGCAGAGGG 3', reverse 5' GCTGGTTTTCACATATGTGC 3') and for its CTRs (forward 5' CCATAGGAACGATCGGTCG 3', reverse 5' CAGCTATATGATATAGTGGTCCG 3'). We cloned PCR products of 240 bp for the block A and 150 to 600 bp for the CTRs representing from 1 to 4 monomers. These fragments were cloned in the pGEM-T vector (Promega) and sequenced to confirm insert specificity. Recombinant plasmids were labeled with digoxigenin 11-dUTP or biotin 11-dUTP by nick translation (Roche Applied Science). Fluorescence *in situ* hybridizations (FISH) were performed as described in Kuhn et al. (2008). Briefly, denaturation of metaphase and polytene chromosomes was carried out in 0.07M NaOH for 3 minutes and 100-200 ng of each probe were hybridized to the chromosomes for 16-20 hours at 37°C in a moist chamber. Slides were washed twice in 2xSSC at 37°C for 5 minutes. DNA fibers were denatured in 70% formamide/2x SSC at 80°C. The slides were analyzed under an Axio Imager A2 epifluorescence microscope equipped with the AxiocamMrm camera (Zeiss). Images were captured with Axiovision (Zeiss) software and edited in Adobe Photoshop.

RNA-Seq analysis and identification of *DINE-TR1* at piRNA clusters

We analyzed publicly available small RNA datasets from the Short Read Archive (SRA) under BioProject GSE22067 produced by Rozhkov et al. (2010). Quality checks and filtering were performed using the FASTX toolkit (Gordon and Hannon 2010) implemented in a local Galaxy instance in Biolinux (Field et al. 2006; Goecks et al. 2010). Short-read mapping against the *Helitron-2_DVir* consensus was performed with Lastz (Harris 2007) also implemented on a local Galaxy instance and excluded reads that mapped with less than 85% identity. Read counts were normalized to one million reads after filtering for abundant degradation transcripts, such as those from tRNAs and rRNAs.

We repeat masked the 20 genomic regions defined as piRNA clusters by Rozhkov et al. (2010) using the CENSOR tool and the *Drosophila* repeat library in Repbase (Kohany et al. 2006). Then, we calculated the contribution of *DINE-TR1* and *Helitrons* in general to each cluster size.

Results

Identification and characterization of *DINE-TR1* in *Drosophila* and other Diptera

After aligning the single most abundant TRs identified in Melters et al. (2013) from the sequenced genomes of 21 *Drosophila* species we realized that in *D. virilis* and *D. elegans* the most abundant TR had a similar size of about 150 bp with high sequence similarity (89%) over a segment of 46 bp. The same 150 bp TR has been found as making short arrays within the *Helitron-2_DVir* of *D. virilis* (Fig. 1B; Abdurashitov et al. 2013). Accordingly, after repeat-masking the TRs in Repbase, we found that the most abundant TR from *D. elegans* also exists as part of a *Helitron* (*Helitron-N1_DEl*; Fig. 1B; Kapitonov and Jurka 2007a).

These two *Helitrons* from *D. virilis* and *D. elegans* belong to the *DINE* group of TEs (Fig. 1A; Kapitonov and Jurka 2007a; Yang and Barbash 2008). We then used the software Tandem Repeats Finder (Benson 1999) to search the Repbase *Helitron* library for elements from other *Drosophila* species that also contained internal TRs similar to those of *D. virilis*. We found that *DINEs* may or may not harbor CTRs and that different and probably unrelated groups of CTRs could be defined based on sequence similarity (data not shown). We named the specific elements with homologous ~150 bp CTRs discussed in this work *DINE-TR1*.

We next surveyed the sequenced genomes of 25 *Drosophila* species with their own *DINE-TR1* consensus from Repbase or with the consensus from the closest related species. The data on the presence or absence of *DINE-TR1* were plotted onto a *Drosophila* plus outgroups phylogeny and showed a discontinuous distribution of *DINE-TR1* across the sampled species (Fig. 2). Entire species subgroups such as *melanogaster* and *pseudoobscura* appeared to be devoid of *DINE-TR1*, whereas in other lineages (e.g. the *virilis-repleta* radiation) *DINE-TR1* presence was patchy (Fig. 2).

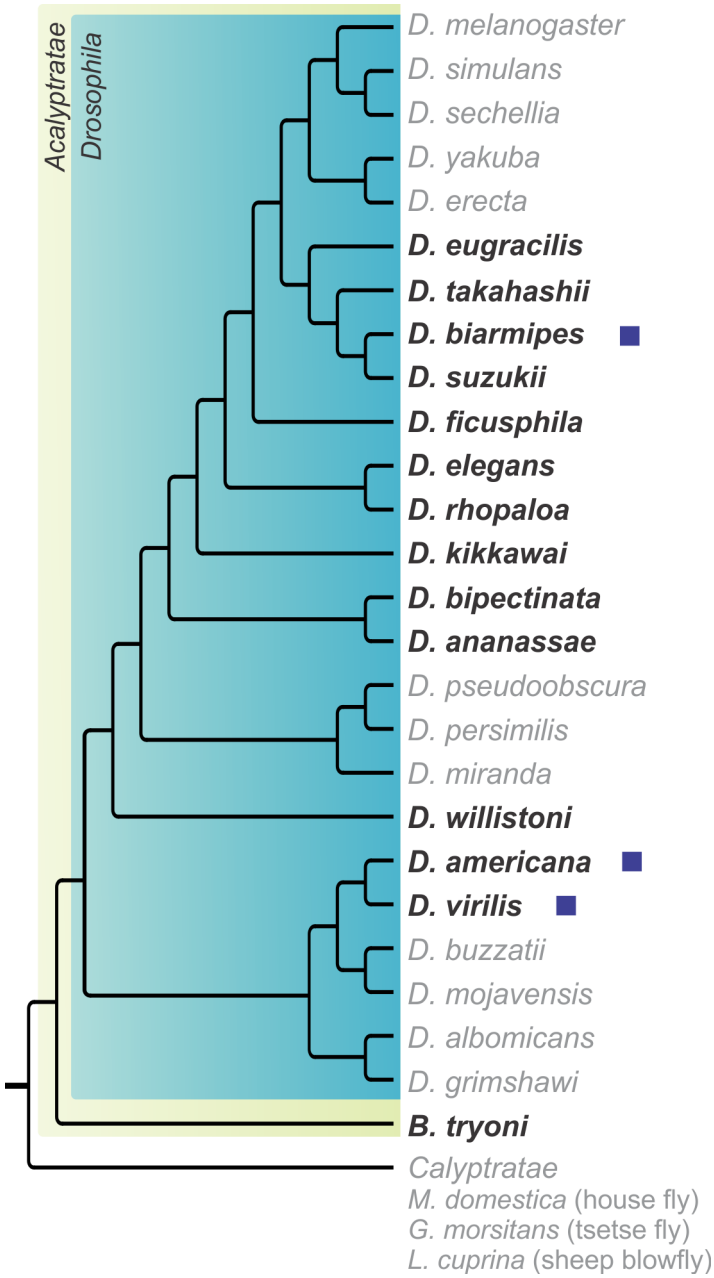


Fig. 2. Phylogeny of Schizophora (Diptera) with representative sequenced species (*Drosophila* phylogenetic relationships were based on Markow 2015 whereas other Diptera species were placed in the tree according to the NCBI Taxonomy Browser classification). Species names in bold indicate the presence of *DINE-TR1* based on genomic analyzes. Blue squares indicate species where *DINE-TR1* CTRs have expanded into abundant satDNA-like arrays.

A BLAST search in the nr/nt (non redundant nucleotide collection) with the *Helitron-2N_DVir* as query and excluding *Drosophila* revealed the presence of *DINE*-like elements in species from several genera within Schizophora (Diptera), among them *Bactrocera*, *Musca*

and *Stomoxys* (Suppl. Table 1). However, in most cases, similarity was restricted to the core region from block A (Fig. 1A). Because there is a limited and uneven availability of sequences from different species in the nr/nt database (which is strongly focused in euchromatic/coding sequences), we advanced our search by using only species with complete sequenced genomes. This approach allowed a more comprehensive search and manual check of the retrieved sequences.

BLAST searches using the *D. virilis DINE-TR1* consensus as a query retrieved several *hits* in all species surveyed. Nevertheless, manual verification of the contigs revealed that only in *Drosophila* and *Bactrocera* the similarity extends over the core region of block A and includes part of the CTRs, indicating that *DINE-TR1* might be restricted to Acalyptratae (Fig. 2). All the three Calyptratae genomes analyzed seem to lack *DINE-TR1* although still possessing other *DINE*-related sequences (Fig. 2; Suppl. Table 1).

To gain insight regarding the discontinuous distribution of *DINE-TR1* inside *Drosophila* we reconstructed a maximum likelihood phylogeny using the entire block A followed by the first CTR from all surveyed species. Although some nodes showed low bootstrap support, the resulting tree topology showed no major incongruence relative to the species tree (Suppl. Fig. 1).

Although the CTRs from *DINE-TR1* have preserved an approximate length of 150 bp in all surveyed species, including in the distantly related *B. tryoni*, high sequence identity among all species was found only in the first ~30 bp of each CTR monomer (88% on average; Fig. 3).

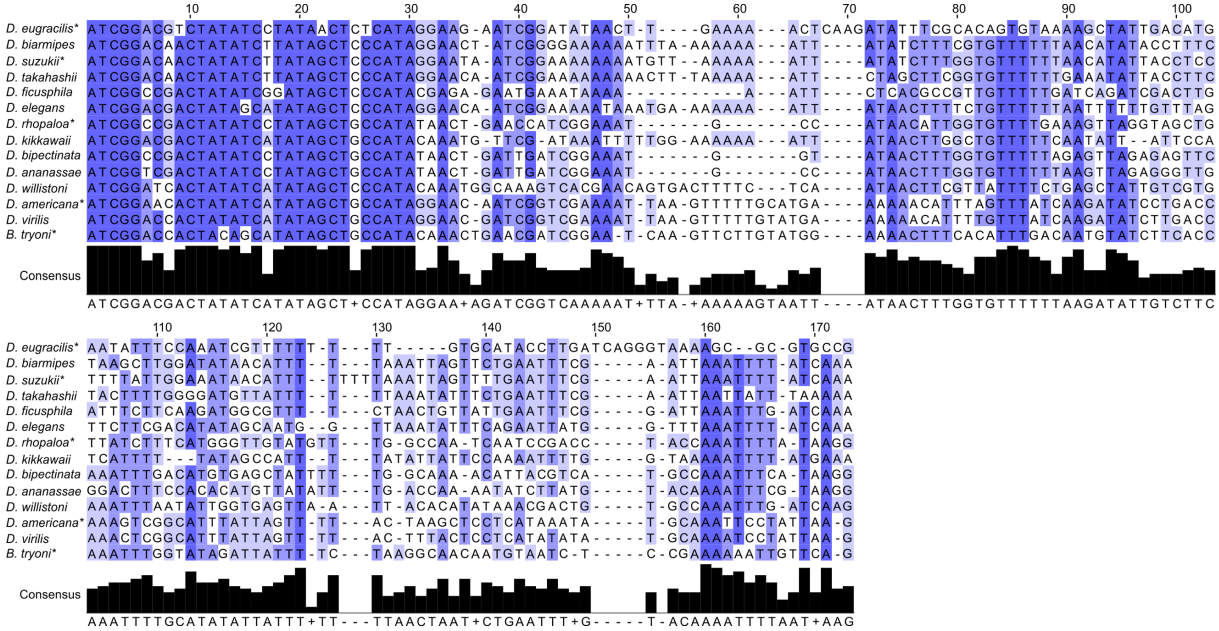


Fig. 3. Multiple sequence alignment (MSA) of the 150 bp CTRs from *DINE-TR1* in several *Drosophila* species and *B. tryoni*. These sequences represent the Repbase *Helitron* consensus for each species. The species with no consensus available are marked with an asterisk and feature the best BLAST hit with the closest species consensus as query. The conservation histogram and consensus sequence for the entire alignment are included bellow the MSA. Dashes represent gaps and the plus symbol represents positions where no majority consensus was reached.

DINE-TR1 with expanded CTRs

Sequence analysis of several contigs of *D. virilis* presenting 150 bp TRs revealed that aside from their organization as short to medium sized arrays inside *DINE-TR1*, these repeats also form larger arrays that cover several Kb (Suppl. Table 3). For example, contigs 0 and 6695 are covered by 150 bp repeats forming arrays of 4376 and 5614 bp, respectively (Suppl. Table 3). Fluorescence *in situ* hybridization (FISH) to extended DNA fibers (fiber FISH) from *D. virilis* using *DINE-TR1* probes specific for the 150 bp TRs and for block A showed intense clustering of *DINE* insertions in some fibers and a marked overabundance of 150 bp TRs in many others (Fig. 4). Altogether, sequence analysis and fiber FISH results suggest that *DINE-TR1* internal CTRs have undergone amplification generating satDNA-like arrays in *D. virilis*.

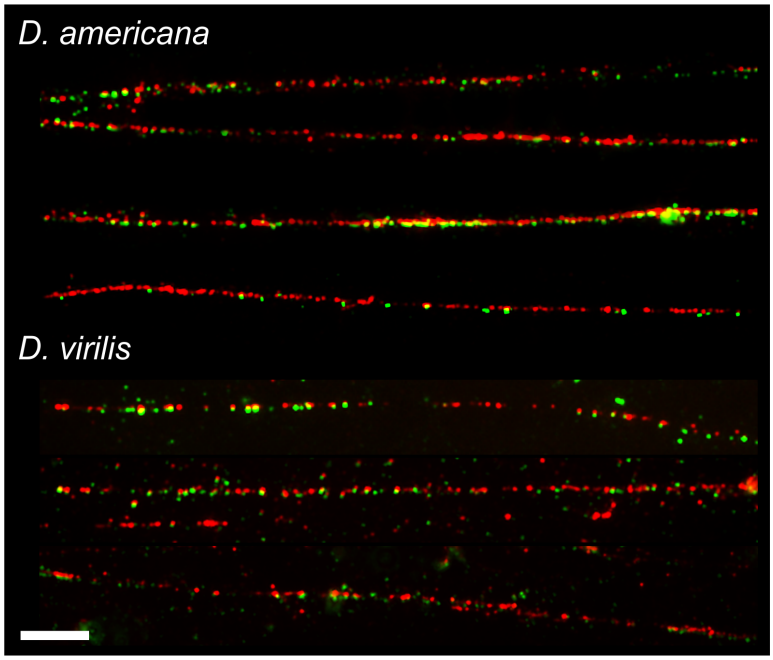


Fig. 4. Fluorescence *in situ* hybridization of *DINE-TR1* block A (green) and the CTRs (red) onto extended DNA fibers from *D. americana* and *D.virilis*. Bar represent 10 kb assuming 10 µm = 29 kb (Schwarzacher and Heslop-Harrison 2001).

The analysis of *DINE-TR1* CTRs in the contigs of *D. elegans* revealed only short arrays (up to six full copies) confined within the *DINE-TR1* structure (Fig. 1C). This result indicates that 150 bp TRs are abundant in the genomes of *D. elegans* due to the high copy number of *DINE-TR1*. Accordingly, we verified that in *D. elegans* all hits from the block A of *DINE* are highly similar to the Repbase consensus (*Helitron-N1_DEl*) for this species (average similarity 99%), possibly indicating a recent transpositional burst.

In order to investigate the amplification status of 150 bp CTRs from *DINE-TR1* in other *Drosophila* species, we isolated the typical CTRs from each species and constructed an artificial array with ten monomers. We then used this array as a query in BLAST searches against the sequenced genome of each species. We found that *D. americana* and *D. biarmipes* also displayed the same satDNA-like arrays of *DINE-TR1* CTRs (Fig. 2; Suppl. Table 3). We also checked the expansion of 150 bp CTRs in *D. americana* using fiber FISH and found a very similar pattern as that described for *D. virilis* (Fig. 4). While *D. americana* is a closely related species to *D. virilis* (belonging to the *virilis* subgroup), *D. biarmipes* is a more distantly related species (*melanogaster* group) (Fig. 2). We further noticed, through *in silico* analysis, that expanded CTRs arrays also exist in other *Drosophila* species, albeit much less abundantly. This was the case for *D. suzukii*, *D. bipectinata* and *Bactrocera tryoni* (Fig. 2).

Besides the amplification suffered by the CTRs of *DINE-TR1*, we also found entire *DINE-TR1* elements arranged in tandem (Fig. 1D). From 80 analyzed cases of nearby insertions, more than half (48) were found separated by 60 bp or less, and 37 were less than 10 bp apart. We found up to 11 *DINE-TR1* tandem repeats (ctg17633; Fig. 1D), to our knowledge the largest number of *Helitron* tandem insertions detected to date. Additionally, we found distinct tandemly organized *Helitrons* (ctg10514, Fig. 1D). This type of insertion was previously reported in maize and, to our knowledge, this is the first reported case in animals (Du et al. 2008).

DINE-TR1 distribution in metaphase and polytene chromosomes of *D. virilis* and *D. americana*

The *D. virilis* karyotype is composed of six acrocentric chromosome pairs (2n=12), with large heterochromatic blocks extending from the centromeres of chromosomes 2, 3, 4, 5 and X and occupying about half of each chromosome. The Y chromosome is entirely heterochromatic and the microchromosome is predominantly euchromatic (Mahan and Beck 1986). *D. americana* shows a derived karyotype by two centromeric fusions (2;3 and X;4) and a similarly distributed but less abundant heterochromatin compared to *D. virilis* (Mahan and Beck 1986; Caletka and McAllister 2004). In order to assess the chromosome distribution of both *DINE-TR1* block A and its CTRs we performed dual-color FISH with specific probes onto the *D. virilis* and *D. americana* metaphase and polytene chromosomes.

In *D. virilis*, the hybridizations revealed a marked enrichment of *DINE-TR1* in the boundaries between the pericentromeric heterochromatin and the euchromatin (i.e. β-heterochromatin) of chromosomes 2, 3, 4, 5 and X (Fig. 5A-B). In *D. americana* we detected signals in similar regions for chromosomes 2, 3, 5 and X, with chromosome 4 only displaying hybridization of the block A probe (Fig. 5C-D). We confirmed that those regions corresponded to the β-heterochromatin through hybridization of the same probes onto polytene chromosomes, which displayed intense co-localizing signals over the entire chromocenter region (Fig. 6). Apart from β-heterochromatin, *DINE-TR1* is abundant in the centromeric region of chromosome 5 and present at a discrete site in the α-heterochromatin of the X chromosome in *D. virilis*, but not in *D. americana*. *DINE-TR1* also covers much of the Y chromosome length in both species (Fig. 5). The microchromosomes showed hybridization signals in both metaphases and polytene chromosomes (Figs. 5 and 6).

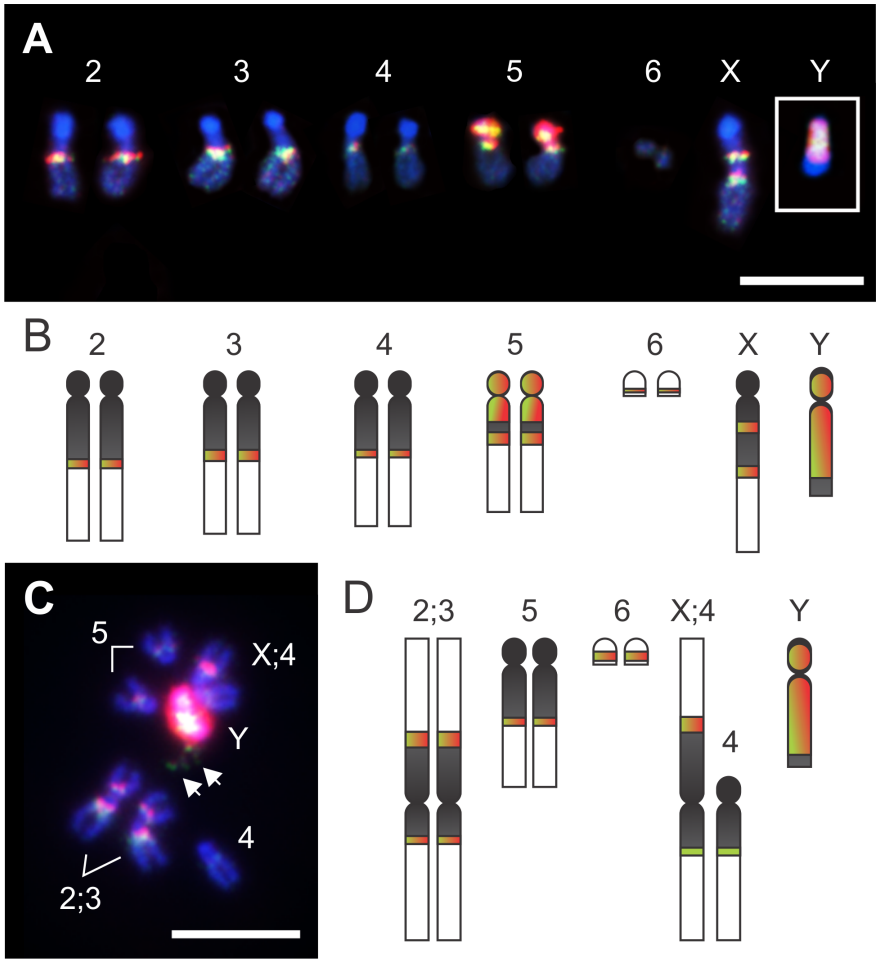


Fig. 5. Fluorescence *in situ* hybridization (FISH) of *DINE-TR1* block-A (green) and 150 bp CTRs (red) onto the metaphase chromosomes of (A) *D. virilis* and (C) *D. americana*. Idiograms of the metaphases and FISH signals are depicted in (B) and (D) with colocalization of the probes represented as a red/green mixed pattern. Black-colored regions represent the constitutive heterochromatin visualized by C-banding (Mahan & Beck 1986). Bars represent 5 µm.

The hybridization of the probes to the polytene chromosomes evidenced the dispersion of *DINE-TR1* at numerous euchromatic loci in all polytene arms, including some telomeric regions (Fig. 6). BLAST searches revealed that *DINE-TR1* is located near or within several genes in *D. virilis* (Suppl. Table 2). It is noteworthy that we found *DINE-TR1* associated with many development-related genes, including several Homeobox genes (Suppl. Table 2).

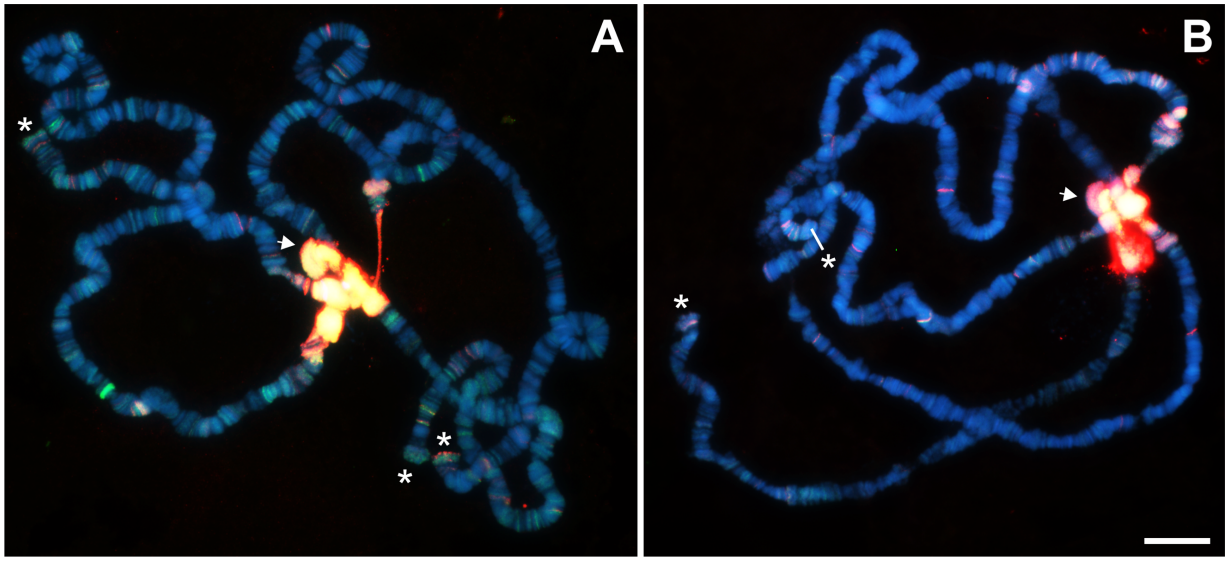


Fig. 6. (A) FISH onto the polytene chromosomes of (A) *D. virilis* and (B) *D. americana* using *DINE-TR1* probes for block A (green) and 150 bp CTRs (red). The chromocenter shows intense hybridization and the small dot chromosome arm is indicated with an arrowhead. Telomeres with hybridization signals are indicated with an asterisk. The bar represent 10 μm.

DINE-TR1 derived small RNAs are abundant in the gonadal tissues of *D. virilis*

The production of piRNAs is thought to occur from clusters of repetitive DNA located at chromatin boundaries (Brennecke et al. 2007). The overall abundance and enrichment of *DINE-TR1* in the β-heterochromatin suggests its possible participation in the piRNA biogenesis. We addressed this issue by mapping public available short-read RNA sequencing data from *D. virilis* (strain 160; Rozhkov et al. 2010) to the *Helitron-2_DVir* consensus sequence. The *Helitron-2_DVir* has been specifically chosen because it represents the full length *Helitron* from the *DINE-TR1* group.

Read counts were calculated for 0-2h embryos and for gonads and carcasses of adult males and females. The results revealed that *Helitron-2_DVir* is almost entirely transcribed, including the CTR region and 5' end (block A). *Helitron-2* small RNA transcripts are relatively abundant in the gonadal tissues from both males and females and almost absent in adult carcasses (Fig. 7A). Interestingly, *Helitron-2_DVir* displays an intermediate transcription level in early embryos (0-2h) relative to adult gonads (Fig. 7A). The mapped reads from embryos and gonads exhibit a medium size of 25,4 nucleotides (nt) while the few mapped reads from adult carcasses display a medium size of 21,5 nt (Fig. 7B). The mapped reads from embryos and ovaries did not show strand bias, whereas reads from male carcass and gonads, and to a lesser degree female carcass, showed abundance of sense over antisense transcripts (Fig. 7C).

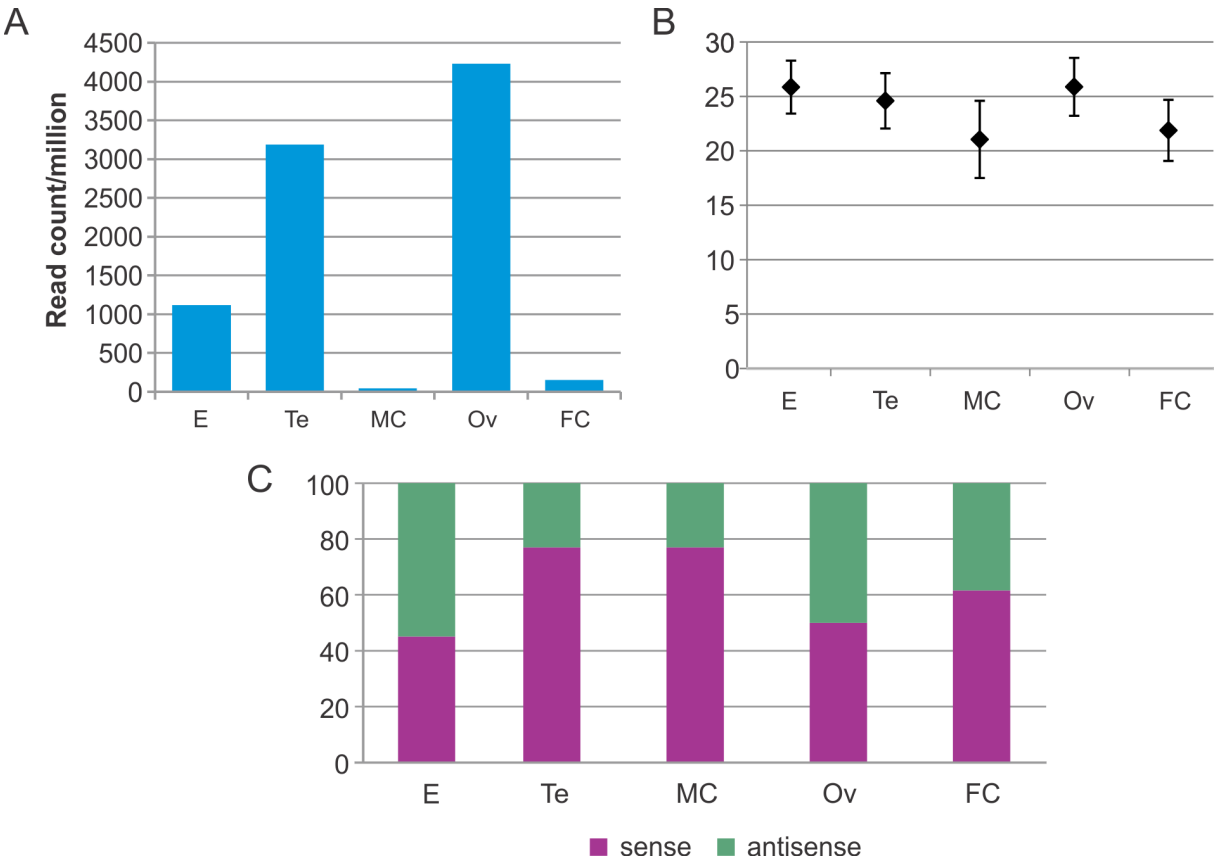


Fig. 7. Characteristics of small RNAs derived from the *DINE-TR1* elements in *D. virilis*. (*A*) Read counts for small RNAs derived from *DINE-TR1* in tissues of *D. virilis* strain 160. Counts were normalized to one million reads. (*B*) Medium size and standard deviation of the small RNAs mapped to *DINE-TR1*. (*C*) Strand bias of *DINE-TR1* transcription. E = 0-2h embryos, Te = testes, MC = male carcass, Ov = ovaries and FC = female carcass.

As a second line of investigation, we repeat masked the genomic regions defined as piRNA clusters by Rozhkov et al. (2010) using the CENSOR tool (Kohany et al. 2006). We found that *DINE-TR1* can be detected in 17 clusters (out of 20) where its abundance range from 0.4 to 9.2% of the total cluster sizes (Suppl. Table 4). Overall, *Helitrons* are the most abundant DNA transposons in these clusters, spanning from 0.5 to 12.1% of the total cluster length (Suppl. Table 4).

Discussion

DINE-TR1 is an ancient group of *Helitrons* from Acalyptratae, Diptera

Helitrons are a poorly understood group of DNA transposons that do not possess typical terminal inverted repeats (TIRs) and are thought to transpose via a rolling circle mechanism (Kapitonov and Jurka 2007b). An interesting and widespread group of *Helitrons* from *Drosophila* are the *DINE-1* elements (Locke et al. 1999). Yang and Barbash (2008) observed that the CTRs of *DINE*s from the sequenced genomes of 12 *Drosophila* species are very variable and do not share interspecies homology except for very closely related species, such as in the ones from the *D. melanogaster* subgroup. Nevertheless, we observed sequence similarity between the CTRs from *DINEs* present in *D. virilis* and *D. elegans*, which indicated that even distant *DINE* elements may have homologous CTRs. This new finding prompted us to analyze this *DINE* group and its CTRs in more detail. Our analyses evidenced the existence of a subset of *DINE* elements (*DINE-TR1*) in Acalyptratae (Diptera). We found *DINE-TR1* in the sequenced genomes of 13 *Drosophila* species (out of 25) and in the Queensland fruit fly *Bactrocera tryoni* but not in out-group species from Calyptratae. This result suggests that *DINE-TR1* was already present in the common ancestor of Acalyptratae, some 72 mya (Gaunt et al. 2002). Interestingly, *DINE-TR1* distribution inside *Drosophila* is patchy (Fig. 2).

The discontinuous distribution of TEs through phylogenies is often explained by means of horizontal transfer (Loreto et al. 2008). However, the phylogenetic reconstruction using *DINE-TR1* sequences did not indicate any major incongruence when compared to the established phylogenetic relationships between these species. In the light of this finding, the patchy distribution of *DINE-TR1* in *Drosophila* could be due to the repeated lineage-specific loss of this element. In fact, Petrov and Hartl (1998) verified that DNA loss is very frequent in *Drosophila*, occurring at a 60 times higher estimated rate than in mammals. This could account for *DINE-TR1* loss especially at an evolutionary time point where its genomic abundance was still low. Alternatively, rapid CTR sequence divergence could also prevent the identification of *DINE-TR1* in some species. In fact, only a small portion of CTRs is conserved between distant species. This indicates an ancient origin for the CTRs of *DINE-TR1* but different evolutionary constraints operating over the monomer. In this context, it is interesting to mention that conserved noncoding blocks (intergenic and intronic) in *Drosophila* are usually small (19 bp on average) and some of them may act as *cis*-regulatory elements (Bergman and Kreitman 2001).

DINE-TR1 as a potential source for satDNA emergence

Our results showed an expansion of the *DINE-TR1* CTRs in *D. virilis*, *D. americana* and *D. biarmipes*, generating satDNA-like arrays. Because *D. virilis* and *D. americana* are both members of the *virilis* species subgroup and share a recent common ancestor dated at ~4 mya (Morales-Hojas et al. 2011), CTR amplification most probably started before the cladogenesis event that separated these two species. On the other hand, the ancient divergence time between *D. biarmipes* and *D. virilis*, at ~62 mya (Tamura et al. 2004), and the lack of a similar pattern of amplification in the other *Drosophila* species (Fig. 2) suggest two independent events of satDNA emergence within *DINE-TR1* in these two lineages.

There is growing evidence of the participation of TEs in the formation of satDNAs, including two reported cases in *D. virilis*. Heikkinen et al. (1995) showed that the pvB370 satDNA share sequence similarity with a TE called pDv and more recently, Dias et al. (2014) showed that a foldback DNA transposon called *Tetris* was involved in the generation of satDNA-like arrays (TIR-220). Herein, we report the participation of a *Helitron* (*DINE-TR1*) in the origin of satDNAs in three *Drosophila* species. To our knowledge, this is the first account on the emergence of satDNAs from preexisting CTRs inside *Helitrons* and also the first report showing the independent emergence of satDNA from the same TE in eukaryotes.

*DINE*s may be involved in the generation of satDNAs through different mechanisms. For example, the *DINE*-related *SGM* sequences generated a major satDNA in *D. guanche* (Miller et al. 2000). However, in this case, most of the element's length became tandemly repeated, similar to what happened to the LINE-1 derived centromeric satDNA of Cetaceans (Kapitonov et al. 1998).

Both the microsatellite and CTR regions of *DINEs* display copy number variation between different insertions as well as between species (Yang and Barbash 2008). In the microsatellite region, slippage replication may be an important mechanism that promotes arrays size variation (Charlesworth et al. 1994). For the CTRs, copy-number variation is possibly related to non-reciprocal DNA exchanges, such as those promoted by unequal crossing over (Charlesworth et al. 1994). Recent examples discuss copy number variation of TE-associated tandem repeats towards satellite DNA emergence (Stavovic and Plohl 2013). Scalvenzi and Pollet (2014) proposed a model for the evolution of TE-derived satellite DNAs as a result of both high recombination and the replicative dispersion of the TEs themselves. This process is expected to result in increase in genome size (Scalvenzi and Pollet 2014). Although these copy number variation mechanisms are ubiquitous, only in *D. virilis*/*D. americana* and *D. biarmipes* CTRs expanded leading to the formation satDNA-like arrays. In other species (e.g. *D. suzukii*, *D. bipectinata* and *B. tryoni*), the amplification of CTRs seems

to be at an earlier stage, or other factors may have been operating to prevent their expansion into large arrays. In any case, the current status of CTR array length must result from the balance between size expansion and reduction mechanisms.

It is important to account for possible assemble errors. In the case of tandem repeats, those errors are generally related to the collapsing of similar reads in the same contig thus shrinking the true array size. In this sense, TE-mediated satDNA emergence could be more frequent than what is currently detectable from NGS genome assemblies.

DINE-TR1 is abundant at chromatin transition zones

In *D. virilis* and *D. americana DINE-TR1* is particularly enriched at transitional β-heterochromatin regions (Figs. 5 and 6). Wasserlauf et al. (2015) recently microdissected the *D. virilis* chromocenter region and generated a DNA library for FISH in the polytene chromosomes of both *D. virilis* and *D. kanekoi* (*virilis* group). Interestingly, *D. kanekoi* also showed intense hybridization signals in its β-heterochromatin, suggesting that *DINE*s already could have colonized this region in the common ancestor of the *virilis* group about 8.9 mya (Morales-Hojas et al. 2011; Wasserlauf et al. 2015). A very similar chromosome distribution was reported for another *DINE*-related element (called *PERI*) present in the *Drosophila buzzatii* species cluster (*repleta* group), that diverged from the *virilis* group more than 20 mya (Kuhn and Heslop-Harrison 2011). These examples may reflect a general feature for *DINE*s.

The β-heterochromatin features both euchromatic and heterochromatic characteristics. It is replicated during polytenization but does not develop into a precise banding pattern, appearing as a loose mass of DNA around the chromocenter (Miklos and Cotsell 1990). This region has been regarded as a "transposon graveyard" because it harbors abundant remnants of ancient TE insertions (Vaury et al. 1989). The clustering of *DINE* insertions at the β-heterochromatin could indicate an insertional preference of these TEs for open chromatin regions and/or a reduced effectiveness of natural selection against the deleterious effects of ectopic recombination upon these sequences (Petrov et al. 2011; see also topic on piRNA clusters below). Additionally, *DINE-TR1* abundance in β-heterochromatin may contribute to define the borders between pericentromeric heterochromatin and euchromatin. In this context, it remains to be investigated whether *DINE-TR1* also act as barrier insulators.

In both *D. americana* and *D. virilis*, we found *DINE-TR1* elements located in the vicinity of the telomeres in some chromosomes (Fig. 5). In *D. melanogaster,* three telomeric-specific non-LTR retroelements, *HeT-A* and *TART* and TAHRE , are involved in telomere maintenance (Villasante et al. 2008). Previous studies showed that the telomeres of *D. virilis*

contain the pvB370 satDNA, and the *TART* and HeT-A retroelements (Biessmann et al. 2000; Casacuberta and Pardue 2003; Villasante et al. 2007). However, the pvB370 satDNA is more likely a telomere associated sequence (TAS) (Casacuberta and Padue 2003; Villasante et al. 2007). *DINE-TR1* could be another TAS in at least some *D. virilis* and *D. americana* chromosomes, defining the borders between euchromatin and telomeric regions.

Locke et al. (1999) used a probe containing the entire *D. melanogaster DINE-1* sequence to assess its distribution in the polytene chromosomes of *D. melanogaster*, *D. simulans* and *D. virilis*. His results suggest that, although being abundant in the dot chromosomes from both *D. melanogaster* and *D. simulans*, *DINE-1* is virtually absent from the *D. virilis* dot. Nevertheless, we did observe hybridization of *DINE-TR1* over the dot in *D. virilis* and *D. americana* (Figs. 5 and 6). This difference could have been caused by the sequence divergence of the probe derived from *D. melanogaster*. In fact, *DINE* elements are abundant in the dots of several *Drosophila* species including *D. erecta*, *D. mojavensis* and *D. yakuba* (Leung et al. 2015).

DINE-TR1 in centromeric DNA

Melters et al. (2013) identified 150 bp TRs as the most abundant TR of the *D. virilis* and *D. elegans* genomes and also likely their major centromeric component. Herein we show that centromeric localization of *DINE-TR1* in *D. virilis* is restricted to chromosomes 5 and Y. This result shows the importance of validating bioinformatic data with cytogenetic tools. In the closely related species *D. americana*, we similarly found *DINE-TR1* covering the Y centromeric region, but not the centromere of chromosome 5. This suggests that either *DINE-TR1* fully colonized the centromere of chromosome 5 only in *D. virilis* or, less probably, it was completely removed from the homologous region in *D. americana* in the last ~4 my. In any case, this illustrates a high rate of evolutionary change of *DINE-TR1* even between closely related species.

It is also worth mentioning that in *D. melanogaster,* β-heterochromatin has been shown to be a hotspot for neocentromere formation under experimental overexpression of centromeric-specific histone H3 (also known as CID) (Olszak et al. 2011). Therefore, one might speculate that the expansion of tandem repeats and TEs from β-heterochromatin to centromeres may contribute for the high rate of centromeric satDNAs turnover observed in *Drosophila* and in many eukaryotes (reviewed by Plohl et al. 2014).

DINE-TR1 is enriched on the Y chromosome

Initial investigations on the *D. virilis* satDNA content revealed the presence of three abundant simple satellites located in the heterochromatin of all autosomes and the X chromosome, but almost absent in the highly heterochromatic Y (Gall et al. 1971). Our results show that a large segment of the Y chromosome of *D. virilis* and of its sister species *D. americana* is covered with *DINE-TR1* copies (Fig. 5). A similar abundance of the *DINE*-related element *PERI* was also found in the Y chromosome of species from the *Drosophila buzzatii* cluster (Kuhn and Heslop-Harrison 2011), which may indicate another general feature of *DINEs*. Some studies indicate a clear correlation between sex chromosomes differentiation and repetitive DNAs accumulation, a process favored by the absence or low frequency of recombination that is typical of these chromosomes (reviewed in Charlesworth et al. 2005). Accordingly, the colonization and expansion of *DINEs* may have been an important event during the process of Y chromosome differentiation in *Drosophila* species. Furthermore, it may also have affected sex-specific gene expression. For example, differences in heterochromatic blocks harboring TEs and other repetitive sequences seem to have been involved in Y-linked regulatory divergence among *D. melanogaster* populations (Lemos et al. 2008). In that sense, heterochromatic blocks could serve as "chromatin sinks" for the binding of transcription factors or chromatin regulators, depleting or redistributing them throughout the genome (Dimitri and Pisano 1989; reviewed in Francisco and Lemos 2014). Such process is thought to be independent of any specific sequence, being a quantitative phenomenon derived from the amount of heterochromatin (Dimitri and Pisano 1989). Interestingly, Brown and Bachtrog (2014) showed that *Drosophila* males have less repressive chromatin modifications in the assembled portions of the genome, which are mostly euchromatic, probably as a result of the Y-derived genome-wide chromatin regulation.

DINE-TR1-derived piRNAs in *D. virilis*

RNA interference (RNAi), or RNA silencing, is a major genomic regulatory mechanism of eukaryotes that recognizes targets by complementarity with small RNAs from three different classes: siRNAs, miRNAs and piRNAs. The interaction of piRNAs with proteins from the Piwi clade (PIWI, AUB and AGO3 in *Drosophila*) is the main genome defense mechanism against transposition events in germ line cells of animals, ensuring stable gametogenesis (Aravin et al. 2007). Nevertheless this class of small RNAs is the least investigated when compared with siRNAs and miRNAs (reviewed in Ghildiyal and Zamore 2009; Siomi et al. 2011). About 90% of *Drosophila* piRNAs can be assigned to TEs, satDNAs, and other repetitive sequences (Brennecke et al. 2007; Yin and Lin 2007; Huang et al. 2013).

DINE-1 is the most abundant transposable element in *Drosophila* (Bergman et al. 2006; Yang et al. 2006; Thomas et al. 2014). Despite some investigations on the piRNA

biogenesis in *D. virilis* (Rozhkov et al. 2010, Le Thomas et al. 2014), the involvement of *DINE-1* elements has not been addressed so far.

DINE copies are heavily accumulated at the β-heterochromatin of *D. virilis* and *D. americana* (Figs. 5 and 6). In *D. melanogaster,* the β-heterochromatin is enriched with fragmented and nested TEs (Vaury et al. 1989; Hoskins et al. 2002). In addition, piRNA clusters have also been shown to map to these regions representing chromatin boundaries (Brennecke et al. 2007; Yamanaka et al. 2014).

The piRNA pathway in *D. melanogaster* is mostly active in gonadal tissues (Brennecke et al. 2007; Brower-Toland et al. 2007; Rozhkov et al. 2010; Le Thomas et al. 2013) and piRNA clusters are generally transcribed from both strands, with no pronounced bias (Brennecke et al. 2007; Rozhkov et al. 2010). The piRNAs have a typical size distribution between 23-29nt; with an average of 25.7, 24.7 and 24.1 nt for Piwi, Aub and Ago3, respectively (Brennecke et al. 2007). We found that small RNA transcripts from *DINE-TR1,* with an average size of 25 nt, are predominantly expressed in *D. virilis* testes and ovaries (Fig. 7). This result strongly points to an active targeting of *DINE-TR1* by the piRNA machinery of *D. virilis*. Interestingly, transcripts from male gonads and carcasses and female carcasses showed strand bias (Fig. 7C). In the case of males, clusters present on the Y chromosome could be skewing piRNA production towards sense strand transcription. Additionally, other classes of small RNAs could be transcribed from *DINE-TR1* in a few loci at low abundances.

When analyzing the genomic regions defined as piRNA clusters by Rozhkov et al. (2010) we found that *DINE-TR1* is present in most of the clusters (Suppl. Table 4). Because there is more than 80 kb of *DINE-TR1* sequences distributed among these clusters, it is plausible to expect that at least some of them are actively transcribed into piRNAs. Altogether, our results strongly suggest the targeting of *DINE*s in *D. virilis* (and probably in other *Drosophila* species) by the piRNA machinery.

DINEs **and chromatin modulation**

It has become clear in the recent years that RNAi pathways are not only essential for germline stability but can also be considered as key factors influencing heterochromatin dynamics (reviewed in Slotkin and Martienssen 2007). For example, Huang et al. (2013) demonstrated that Piwi-piRNA complexes interact with chromatin factors such as HP1a (Heterochromatin Protein 1a) and HMTs (Histone Methyltransferases) guiding them to specific locations and promoting chromatin changes in a genome-wide scale.

We found that the proportion of *DINE*-derived RNAs in 0-2h *D. virilis* embryos is intermediate between the values found in gonads and carcasses (Fig. 7A). Because these

embryos have not fully onset transcription (Vlassova et al. 1991; Pritchard and Schubiger 1996), their small RNAs and Piwi proteins are essentially the same of the maternal germ cells (Harris and Macdonald 2001; Megosh et al. 2006; Brennecke et al. 2008; Le Thomas et al. 2014). A similar scenario was found for *D. melanogaster* (Brennecke et al. 2008). At early embryogenesis, the maternally-inherited piRNAs and piwi proteins could be leading elements in the process of defining heterochromatic domains (Sentmanat et al. 2013). Heterochromatin formation is triggered in embryo cells around 2h old, coinciding with the first signs of transcription in the embryo cells themselves (Vlassova et al. 1991). The smaller proportion of *DINE*-derived RNAs in 0-2h embryos could be the result of differences between somatic and germ cell transcripts from the ovaries (with a larger proportion of *DINE-TR1* transcripts in somatic cells), the normal depletion of maternally-inherited piRNAs during early development, or both.

Our description of *DINE-TR1* association with several *D. virilis* genes agrees with the previous finding that *DINE*s are frequently located within introns and flanking coding regions in several *Drosophila* species (Yang and Barbash 2008). Interestingly, the same study reports that *D. virilis* has the highest number of intronic insertions (1,104) among the 12 *Drosophila* species analyzed. It is possible that small RNAs interact with some of these *DINE* elements, establishing local chromatin modifications and affecting the regulation of genes. For example, in the *Arabidopsis* accession Landsberg erecta (Ler), the FLC gene (a key factor in flowering pathways) has a Mutator-like transposon insertion in the first intron, which is responsible for its low expression (Gazzani et al. 2003; Michaels et al. 2003). Furthermore, it has been demonstrated that this TE insertion is involved in siRNA-mediated silencing by forming a heterochromatin "island" restricted to the element and its vicinity (Liu et al. 2004). More recently, intronic insertions of *Helitrons* on *Arabidopsis* and rice genes have been shown to be the main targets for heterochromatin establishment (Saze et al. 2013).

In *D. melanogaster*, TE insertions next to genes have been associated with changes in the chromatin state such as the di- and trimethylations of lysine 9 from histone H3 (H3K9me), and HP1a assembly; which are typical heterochromatic marks. These chromatin changes are likely piRNA-mediated and result in lower expression of the nearby genes (Sentmanat and Elgin 2012; Lee 2015). Those recent findings indicate that piRNA-mediated TE silencing is not restricted to heterochromatic TE insertions and that abundant TEs such as *DINE-TR1* in *D. virilis* could have a huge impact in both chromatin modulation and gene regulation.

Acknowledgements

We would like to thank Claudia Carareto (Universidade Estadual Júlio Mesquita Filho, Brazil) and Jorge Vieira (Instituto de Biologia Molecular e Celular, Portugal) for providing the *D. virilis* and *D. americana* stocks, respectively. We are also grateful for the anonymous reviewers' comments on the manuscript. This work was supported by grants from “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG) (Proc: APQ-01563-14), “Programa Institucional de Auxílio à Pesquisa de Doutores Recém-Contratados da Universidade Federal de Minas Gerais”, “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq) and a doctoral fellowship from "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) to GD.

References

Abdurashitov MA, Gonchar DA, Chernukhin VA, Tomilov VN, Tomilova JE, Schostak NG, Zatsepina OG, Zelentsova ES, Evgen'ev MB, Degtyarev SK (2013) Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC Genomics 14(1):771

Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318(5851):761–764

Ashburner M (1989) *Drosophila*. A laboratory handbook. Cold Spring Harbor Laboratory Press

Baimai V (1977) Chromosomal polymorphisms of constitutive heterochromatin and inversions in *Drosophila*. Genetics 85(1):85–93

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 27(2):573–580

Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res 1(8):1335–1345

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. Genome Biol 7(11):R112

Biessmann H, Zurovcova M, Yao JG, Lozovskaya E, Walter MF (2000) A telomeric satellite in *Drosophila virilis* and its sibling species. Chromosoma 109(6):372–380

Bosco G, Campbell P, Leiva-Neto JT, Markow TA (2007) Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177(3):1277–1290

Brajkovic J, Feliciello I, Bruvo-MadWaric B, Ugarkovic D (2012) Satellite DNA-Like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 (Bethesda) 2:931–941

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128(6):1089–1103

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322(5906):1387–1392

Brower-Toland B, Findley SD, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin SCR, Lin H (2007) *Drosophila* PIWI associates with chromatin and interacts directly with HP1a. Genes Dev 21(18):2300–2311

Brown EJ, Bachtrog D (2014) The chromatin landscape of *Drosophila*: comparisons between species, sexes, and chromosomes. Genome Res 24:1125–1137

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. Annu Rev Genomics Hum Genet 11:291–316

Carareto CM, Hernandez EH, Vieira C (2014) Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling *Drosophila* species. Gene 537(1):93–99

Casacuberta E, Pardue ML (2003) Transposon telomeres are widely distributed in the *Drosophila* genus: *TART* elements in the *virilis* group. Proc Natl Acad Sci USA 100(6):3363–3368

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–220

Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118–128

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature methods 9(8):772–772

de Wit E, Greil F, van Steensel B (2005) Genome-wide HP1 binding in *Drosophila*: developmental plasticity and genomic targeting signals. Genome Res 15:1265–1273

Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GCS (2014) Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. Genome Biol Evol 6(6):1302–1313

Dimitri P, Pisano C (1989) Position effect variegation in *Drosophila melanogaster*: relationship between suppression effect and the amount of Y chromosome. Genetics 122(4):793–800

Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, Tally JF, Notredame C (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res gkr245:W13-W17

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

Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9(1):51

Ellison CE, Bachtrog D (2013) Dosage compensation via transposable element mediated rewiring of a regulatory network. Science 342(6160):846–850

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković Đ (2015) Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. Genome Biol Evol 7(1):228–239

Ferree PM, Barbash DA (2009) Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in *Drosophila*. PLoS Biol 7(10):e1000234. doi:10.1371/journal.pbio.1000234

Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M (2006) Open software for biologists: from famine to feast. Nat Biotechnol 24(7):801–804

Francisco FO, Lemos B (2014) How Do Y-Chromosomes Modulate Genome-Wide Epigenetic States: Genome Folding, Chromatin Sinks, and Gene Expression. J Genomics 2:94–103

Fonseca NA, Morales-Hojas R, Reis M, Rocha H, Vieira CP, Nolte V, Schlötterer C, Vieira J (2013) *Drosophila americana* as a model species for comparative studies on the molecular basis of phenotypic variation. Genome biology and evolution 5(4):661–679

Gall JG, Cohen EH, Polan ML (1971) Repetitive DNA sequences in *Drosophila*. Chromosoma 33(3):319–344

Gatti M, Pimpinelli S, Santini G (1976) Characterization of *Drosophila* Heterochromatin. Chromosoma (Berl.) 57:351–375

Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol 19(5):748–761

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Physiol 132:1107–1114

Gilchrist AS, Shearman DC, Frommer M, Raphael KA, Deshpande NP, Wilkins MR, Sherwin WB, Sved JA (2014) The draft genome of the pest tephritid fruit fly *Bactrocera tryoni*: resources for the genomic analysis of hybridising species. BMC Genomics 15(1):1153

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86

Gordon A, Hannon GJ (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab.cshl.edu/fastx_toolkit

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

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455:1193–1197

Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nature Rev Genet 10(2):94–108

Guillén Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, Casillas S, Ràmia M, Egea R, Negre B, Mir G, Camps J, Moncunill V, Ruiz-Ruano FJ, Cabrero J, de Lima LG, Dias GB, Ruiz JC, Kapusta A, Garcia-Mas J, Gut M, Gut IG, Torrents D, Camacho JP, Kuhn GC, Feschotte C, Clark AG, Betrán E, Barbadilla A, Ruiz A (2014) Genomics of ecological adaptation in cactophilic *Drosophila*. Genome Biol Evol evu291

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology. 52(5):696–704

Harris AN, Macdonald PM (2001) Aubergine encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. Development 128(14):2823–2832

Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University

Heikkinen E, Launonen V, Müller E, Bachmann L (1995) The pvB370 BamHI Satellite DNA Family of the *Drosophila virilis* Group and Its Evolutionary Relation to Mobile Dispersed Genetic pDv Elements. J Mol Evol 41:604–614

Heslop-Harrison JS, Schwarzacher T (2011) Organization of the plant genome in chromosomes. Plant Journal 66:18–33

Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, Yasuhara JC, Wakimoto BT, Myers EW, Celniker SE, Rubin GM, Karpen GH (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biol 3(12):research0085

Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H (2013) A major epigenetic programming mechanism guided by piRNAs. Dev Cell 24(5):502–516

International *Glossina* Genome Initiative (2014) Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. Science 344(6182):380–386

Jordan IK, Miller WJ (2008) Genome Defense Against Transposable Elements and the Origins of Regulatory RNA. In: Lankenau DH, Volff JN, eds. Transposons and the Dynamic Genome. Springer-Verlag Berlin Heidelbarg, pp 77–94

Junier T, Pagni M (2000) Dotlet: diagonal plots in a web browser. Bioinformatics 16(2):178–179

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1-4):462–467

Kalmykova AI, Klenov MS, Gvozdev VA (2005) Argonaute protein PIWI controls mobilization of retrotransposons in the *Drosophila* male germline. Nucleic Acids Res 33(6):2052–2059

Kapitonov VV, Holmquist GP, Jurka J (1998) L1 repeat is a basic unit of heterochromatin satellites in cetaceans. Molecular biology and evolution, 15(5):611–612

Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719

Kapitonov VV, Jurka J (2007) *Helitrons* in fruit flies. Repbase Reports 7(3):130–130

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

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

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC bioinformatics 7(1):474

Kuhn GCS, Franco FF, Manfrin MH, Moreira-Filho O, Sene FM (2008) Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single pair of microchromosomes in species from the *Drosophila buzzatii* cluster. Chromosome Res 16:307–324

Kuhn GC, Heslop-Harrison JS (2011) Characterization and genomic organization of PERI, a repetitive DNA in the *Drosophila buzzatii* cluster related to DINE-1 transposable elements and highly abundant in the sex chromosomes. Cytogenet Genome Res 132:79–88

Lee YCG (2015) The Role of piRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in *Drosophila melanogaster*. PLOS Genet 11(6), e1005269

Lemos B, Araripe LO, Hartl DL (2008) Polymorphic Y Chromosomes Harbor Cryptic Variation with Manifold Functional Consequences. Science 319:91–93

Le Thomas A, Marinov GK, Aravin AA (2014) A Transgenerational Process Defines piRNA Biogenesis in *Drosophila virilis*. Cell Reports 8:1617–1623

Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in *Arabidopsis*. Genes Dev 18(23):2873–2878

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

Loreto ELS, Carareto CMA, Capy P (2008) Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100(6):545–554

Macas J, Koblížková A, Navrátilová A, Neumann P (2009) Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448:198-206

Mahan JT, Beck ML (1986) Heterochromatin in mitotic chromosomes of the *Virilis* species group of *Drosophila*. Genetica 68:113-118

Malik HS, Henikoff S (2009) Major Evolutionary Transitions in Centromere Complexity. Cell 138(6):1067–1082

Megosh HB, Cox DN, Campbell C, Lin H (2006) The role of PIWI and the miRNA machinery in *Drosophila* germline determination. Curr Biol 16:1884–1894

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, Garcia JF, DeRisi JL, Smith T, Tobias C, Ross-Ibarra J, Korf I, Chan SW (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14:R10

Menon DU, Coarfa C, Xiao W, Gunaratne PH, Meller VH (2014) siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. Proc Natl Acad Sci USA 111(46):16460–16465

Michaels SD, He Y, Scortecci KC, Amasino RM (2003) Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. Proc Natl Acad Sci USA 100(17):10102–10107

Miklos GLG, Cotsell JN (1990) Chromosome structure at interfaces between major chromatin types: Alpha-and Beta-heterochromatin. Bioessays 12(1):1–6

Miller WJ, Nagel A, Bachmann J, Bachmann L (2000) Evolutionary Dynamics of the SGM Transposon Family in the *Drosophila obscura* Species Group. Mol Biol Evol 17(11):1597–1609

Morales-Hojas R, Reis M, Vieira CP, Vieira J (2011) Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Mol Phylogenet Evol 60(2):249–258

Olszak AM, van Essen D, Pereira AJ, Diehl S, Manke T, Maiato H, Saccani S, Heun P (2011) Heterochromatin boundaries are hotspots for de novo kinetochore formation. Nature Cell Biol 13(7):799–808

Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D, Dekker T, Wrobel N, Viola R, Pertot I, Cavalieri D, Blaxter M, Anfora G, Rota-Stabelli O (2013) Linking genomics and ecology to investigate the complex evolution of an invasive *Drosophila* pest. Genome biology and evolution 5(4):745–757

Petrov DA, Hartl DL (1998) High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol Biol Evol 15(3):293–302

Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J (2011) Population genomics of transposable elements in *Drosophila melanogaster*. Mol Biol Evol 28(5):1633–1644

Plohl M, Meštrović N, Mravinac B (2014) Centromere identity from the DNA point of view. Chromosoma 123:313–325

Pritchard DK, Schubiger G (1996) Activation of transcription in *Drosophila* embryos is a gradual process mediated by the nucleocytoplasmic ratio. Genes Dev 10(9):1131–1142

Rošić S, Köhler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J. Cell Biol 207(3):335–349

Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR, Hannon GJ, Evgen'ev MB (2010) Small RNA-based silencing strategies for transposons in the process of invading *Drosophila* species. Rna, 16(8):1634–1645

Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC (2006) Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. Genes Dev 20:2214–2222

Satovic E, Plohl M (2013) Tandem repeat-containing MITEs in the clam *Donax trunculus*. Genome Biol Evol 5(12):2549–2559

Saze H, Kitayama J, Takashima K, Miura S, Harukawa Y, Ito T, Kakutani T (2013) Mechanism for full-length RNA processing of Arabidopsis genes containing intragenic heterochromatin. Nat Commun 4:2301

Scalvenzi T, Pollet N (2014) Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA. Mol Phylogenet Evol 81:1–9

Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA, Li M, Meisel RP, Minx P, Murphy TD, Nelson DR, Reid WR, Rinkevich FD, Robertson HM, Sackton TB, Sattelle DB, Thibaud-Nissen F, Tomlinson C, van de Zande L, Walden KKO, Wilson RK, Liu N (2014) Genome of the house fly, *Musca domestica L.*, a global vector of diseases with adaptations to a septic environment. Genome Biol 15(10):466

Sentmanat MF, Elgin SC (2012) Ectopic assembly of heterochromatin in *Drosophila melanogaster* triggered by transposable elements. Proc Natl Acad Sci USA 109(35):14104–14109

Sentmanat M, Wang SH, Elgin SCR (2013) Targeting heterochromatin formation to transposable elements in *Drosophila*: Potential roles of the piRNA system. Biochemistry (Moscow) 78(6):562–571

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Rev Genet 8(4):272–285

Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nature Rev Mol Cell Biol 12:246–258

Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol 21(1):36–44

Thomas J, Vadnagara K, Pritham E (2014) DINE-1, the highest copy number repeats in *Drosophila* melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). Mobile DNA 5:18

Ugarkovic D (2009) Centromere-Competent DNA: Structure and Evolution. In: Ugarkovic D, ed. Centromere: Structure and Evolution. Springer-Verlag Berlin Heidelberg. pp. 53–76

Vaury C, Bucheton A, Pelisson A (1989) The b heterochromatic sequences flanking the I elements are themselves defective transposable elements. Chromosoma 98:215–224

Villasante A, Abad JP, Planelló R, Méndez-Lago M, Celniker SE, de Pablos B (2007) *Drosophila* telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. 17(12):1909–1918

Villasante A, de Pablos B, Méndez-Lago M, Abad JP (2008). Telomere maintenance in *Drosophila*: rapid transposon evolution at chromosome ends. Cell Cycle 15;7(14):2134–2138

Vermaak D, Malik HS (2009) Multiple roles for heterochromatin protein 1 genes in *Drosophila*. Annu Rev Genet 43:467–492

Vlassova IE, Graphodatsky AS, Belyaeva ES, Zhimulev IF (1991) Constitutive heterochromatin in early embryogenesis of *Drosophila melanogaster*. Mol Gen Genet 229(2):316–318

Volpe T, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen R (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297(5588):1833–1837

Wallrath LL, Vitalini MW, Elgin SCR (2014) Heterochromatin: A Critical Part of the Genome. In: Workman JL, Abmayr SM, eds. Fundamentals of Chromatin. Springer New York, pp 529–552

Wasserlauf I, Usov K, Artemov G, Anan'ina T, Stegniy V (2015) Specific features in linear and spatial organizations of pericentromeric heterochromatin regions in polytene chromosomes of the closely related species *Drosophila virilis* and *D. kanekoi* (Diptera: Drosophilidae). Genetica 1-12.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191

Yamanaka S, Siomi MC, Siomi H (2014) piRNA clusters and open chromatin structure. Mobile DNA 5(1):22

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

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

Yin H, Lin H (2007) An epigenetic activation role of Piwi and a Piwi associated piRNA in *Drosophila melanogaster*. Nature 450:304–308

Capítulo 2

Helitrons in *Drosophila*: chromatin modulation and tandem insertions

Este capítulo é composto de um artigo teórico, no formato de comentário, publicado na revista *Mobile Genetic Elements*. Neste artigo expandimos a discussão sobre o impacto da heterocromatina formada através de piRNAs sobre sequências derivadas de TEs. Também discutimos a frequente integração de *Helitrons* em tandem e os modelos de transposição que podem explicar este fenômeno.

Helitrons in *Drosophila*: chromatin modulation and tandem insertions

Guilherme B. Dias, Pedro Heringer, Gustavo C. S. Kuhn
Departamento de Biologia Geral; Universidade Federal de Minas Gerais; Belo Horizonte, MG, Brasil.

Commentary to: Dias GB, Heringer P, Svartman M, Kuhn GCS. *Helitrons* shaping the genomic architecture of *Drosophila*: enrichment of *DINE-TR1* in α-and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosome Res 2015; PMID: 26408292; (3):597-613; http://dx.doi.org/10.1007/s10577-015-9480-x

Keywords: β-heterochromatin, piRNA, chromatin sink, gene regulation, *DINE-1*

List of abbreviations:

DINE: *Drosophila* INterspersed Element
HP1: Heterochromatin Protein 1
My: million years
RC: rolling-circle
Rep: replication initiator protein
TE: transposable element
TI: tandem insertion
TR: tandem repeat

Abstract

Although *Helitrons* were discovered 15 years ago, they still represent an elusive group of transposable elements (TEs). They are thought to transpose via a rolling-circle mechanism but no transposition assay has been conducted yet. We have recently characterized a group of *Helitrons* in *Drosophila*, named *DINE-TR1*, that display interesting features, including a pronounced enrichment at β-heterochromatin, multiple tandem insertions (TIs) of the entire TE, and that experienced at least two independent events of expansion of its internal tandem repeats (TRs) in distant *Drosophila* lineages. Here we discuss two aspects of TE dynamics showcased by the *DINE-TR1 Helitrons:* (i) the general evolutionary impact of piRNA-guided heterochromatin formation via TE-derived TR expansion and (ii) the possible mechanisms that could account to the recurrent TIs of *Helitrons*.

Introduction

Eukaryotic genomes are notably rich in several types of repetitive sequences. Repetitive DNA, although mostly parasitic in nature, have been fundamental in shaping these genomes and generated many beneficial side effects by means of exaptation (Kidwell and Lisch 2000; Shapiro and von Sternberg 2005; Elliott et al. 2014). The most abundant type of repetitive DNA are the transposable elements (TEs), which are stretches of DNA able to move between loci in the host genome and make copies of themselves in the process (Gregory 2005). In spite of the large number of studies conducted on these elements over the last decades, TE abundance and diversity is so massive that we still do not fully appreciate their impact on genome evolution.

One particularly elusive group of TEs are the *Helitrons*, DNA transposons thought to mobilize by a mechanism similar to the rolling circle replication of some plasmids, single stranded DNA virus and bacterial transposons (Kapitonov and Jurka 2001). These TEs were discovered 15 years ago in *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Oryza sativa*, and they are now known to be present in a wide phylogenetic range, from fungi to mammals (Kapitonov and Jurka 2001; Kapitonov and Jurka 2007; Thomas and Pritham 2015).

Despite many advances in *Helitron* research over the past years, data are still lacking for their mode of transposition, genome organization, and chromosome distribution, for most studied species. In the genus *Drosophila*, for example, the metaphase chromosome distribution of the prolific *DINE-1 Helitrons* (Locke et al. 1999) is not known for most species, including *Drosophila melanogaster*.

In a recent work, we improved characterization of *Helitrons* from the *DINE-1* group in two *Drosophila* species from the *virilis* group, *D. virilis* and *D. americana*, and defined a subgroup of *DINEs* present in Acalyptratae (Diptera) that we called *DINE-TR1* (Dias et al. 2015). This group is characterized by structural and sequence features, including the presence of internal tandem repeats (TRs) of ~150bp which share a conserved region of 30-40bp in their 5' end, suggesting that this motif may exhibit functionality. In *D. virilis* and *D. americana, DINE-TR1* is located in many euchromatic loci but is particularly enriched in chromatin/heterochromatin boundaries (β-heterochromatin) and in the Y chromosome. In *D. virilis*, *DINE-TR1* further colonized the centromeric region of chromosome 5 (Muller element C). Small RNAs matching *DINE-TR1* are ~25nt in length, abundant in 0-2h embryos and adult gonads, but almost absent in adult carcasses, suggesting that *DINE-TR1* copies are targeted for silencing by the piRNA pathway in *D. virilis* (Dias et al. 2015).

Surprisingly, we detected two independent events of TR amplification from within *DINE-TR1* that occurred in *Drosophila* lineages that diverged over 60 My ago (*virilis* subgroup and *D. biarmipes*). A few other *Drosophila* species and the Queensland fruit fly *Bactrocera tryoni* also showed signs of incipient expansion of *DINE-TR1* internal TRs. These results point to *DINE-TR1* as a recurrent source of satellite DNAs in Acalyptratae.

In this commentary, we focus on two aspects of TE dynamics showcased by the *DINE-TR1 Helitrons*. First, we discuss the evolutionary impact of TR expansion within TEs that are targeted by the piRNA pathway, such as *DINE-TR1*, to gene and genome regulation and evolution. Second, we draw attention to the numerous *DINE-TR1* tandem insertions (TIs) we detected in the *D. virilis* genome assembly, with up to 11 sequential elements. We discuss the literature concerning *Helitron* TIs and highlight some unresolved questions regarding *Helitron* transposition.

piRNA-targeted TEs as sources of TR expansion and heterochromatin formation

There are now several reports on the expansion of TRs from within different classes of TEs (reviewed in Meštrović et al. 2015). The occurrence of internal TRs seems to be a relatively common feature for many TEs, although the functional role of these repeats (if any) is mostly unknown (Macas et al. 2009). For TEs belonging to the terminal inverted repeat (TIR) order it has been proposed that the increased number of internal TRs containing transposase binding sites could improve the specificity of recognition by the transposase leading to a more efficient mobilization (Marzo et al. 2013). In fact, the presence of multiple transposase binding sites improves *in vitro* transposition of the TE *Sleeping Beauty* in HeLa cells (Zayed et al. 2004). For retroelements from the large Ty3/Gypsy family, it has been proposed that TRs at the 3' end could act as transposition termination signals by forming stable stem loops (Martínez-Izquierdo et al. 1997). Similar examples of TRs residing within *Helitrons* have been reported (e.g. Yang and Barbash 2008; Kuhn et al. 2011, Sătović and Plohl 2013), although no explanation regarding their functional significance was evoked yet.

Regardless of their involvement in TE movement itself, the expansion/contraction of internal TRs can impact genome structure in several ways. It has been recently shown that TEs induce local heterochromatin formation via deposition of H3K9me and such epigenetic change may affect the expression of nearby genes (Lee 2015).

Summing to the local effects of repeats on gene regulation, the overall dosage of heterochromatin was shown to affect position effect variegation (PEV) and viability in

Drosophila (Dimitri and Pisano 1989; Berloco et al. 2014). This hypothesis is known as the "chromatin sink" hypothesis, and accounts for the ability of chromatin to recruit modulator proteins in such a way that increases or decreases in heterochromatin dosage, would lead to depletion or overabundance, respectively, of these modulators in other regions of the genome (Dimitri and Pisano 1989).

In light of this knowledge regarding the epigenetic effects of repetitive sequences, it is possible to advance two consequences of TR expansion from piRNA targeted TEs. They might act as "tuning knobs" of gene expression in euchromatin and/or as factories of satellite DNAs that influence both heterochromatin turnover and genome modulation (Fig. 1A-B).

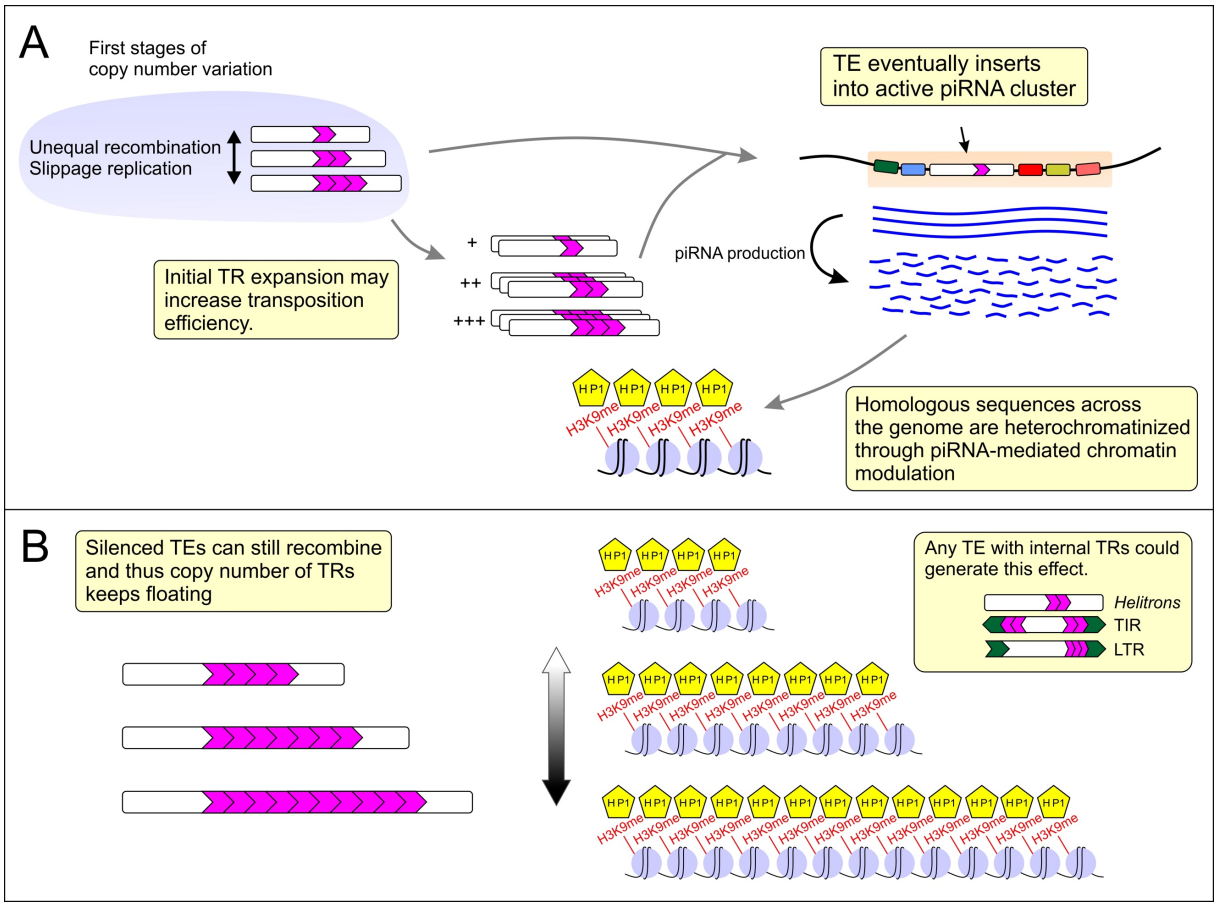


Figure 1. General layout for transposable element (TE)-derived tandem repeat (TR) expansion and heterochromatin formation. (A) At some point in the TE life cycle one of its copies inserts into an active piRNA cluster and then serves as a template for the generation of complementary piRNAs that silence homologous sequences in the genome via heterochromatin formation. (B) The heterochromatinized TE copies harboring internal TRs

are prone to suffer unequal recombination. This TR concertina generates variation in the size of heterochromatin blocks what may in turn affect gene expression.

The TRs within TEs may expand or contract, and this variation is likely the result of either unequal recombination or slippage replication (Charlesworth et al. 1994, Scalvenzi and Pollet 2014). The initial variation in array size may fluctuate by drift or proceed towards expansion in the case that additional TRs increase transposition efficiency (Marzo et al. 2013; Meštrović et al. 2015). Constant transpositions will eventually result in the insertion of a TE copy within an active piRNA cluster. Such event ensues production of piRNAs that will recognize and silence the copies of this TE, reducing or stopping further transpositions (Fig. 1A). The silencing is thought to occur at both transcriptional (via H3K9me and HP1; Le Thomas et al. 2013) and post-transcriptional levels (processing of TE transcripts by PIWI-clade proteins; Brennecke et al. 2007). Although unable to transpose, the dispersed silenced copies of the TE can still be subject to unequal exchange mechanisms such as unequal recombination. In this context, expansion and contractions of internal TRs coupled with piRNA targeting now result in variations in the size of local heterochromatin blocks, acting like a heterochromatin concertina (Fig. 1B).

The heterochromatin formation via piRNA-targeted TEs could have a great impact in driving the turnover of centromeric sequences, since new residents of the centromere must keep the heterochromatic state of this region (Henikoff et al. 2001; Plohl 2014). In plants, a boom-bust process was evoked to explain the radical shifts in sequence composition of homologous centromere regions before a favorable repeat becomes fixed (Zhang et al. 2014). This model could account for the quick colonization of the centromere of chromosome 5 in *D. virilis* by *DINE-TR1*, which is absent in the same region in the close species *D. americana* (Dias et al. 2015). This turnover was likely achieved by both *DINE-TR1* replicative transposition and the heterochromatin formation via TR expansion. Such rapid changes in heterochromatin content may influence genome stability and contribute to species divergence (Kuhn et al. 2008; Ferree and Barbash 2009). Thus, TRs derived from TEs targeted by piRNAs could be a recurrent source for heterochromatin turnover.

In the case where TEs with internal TRs are close to or within genes, the concertina-like array size variation of silenced TRs could modulate gene expression and be in turn subject to selection (King et al. 1997). In *D. virilis*, for example, there are over 1,000 intronic insertions of *DINE-1* (Yang and Barbash 2008) and many of them represent *DINE-TR1* elements. The tuning of *DINE-TR1* internal repeat array length by natural selection is

arguably more dynamic than the mere presence/absence of TEs close to genes. In fact, the epigenetic effect of TEs on gene expression was associated with the number and length of TEs, distance from the gene and also the likelihood that these TEs were targeted by the piRNA pathway (Lee 2015). The same mechanisms of TE silencing are assumed to act upon TE-derived TRs as long as they are also present in the piRNA generating clusters. In this context, it would be interesting to investigate the extent of TR array size variation of *DINE-TR1* associated with genes and in "gene desert" regions. The size variation of TE-derived TRs is also expected to contribute for the bulk amount of heterochromatin, which itself has regulatory properties (Zuckerkandl 1974; Berloco et al. 2014).

In conclusion, the expansion/contraction of TRs within piRNA targeted TEs could impact genome regulation at several levels, according to both the chromatin sink hypothesis and the local regulatory effect exerted by repeats. This could be an important process in tuning global genome regulation and also a step towards population divergence.

Tandem insertions of *Helitrons*: a still unexplained phenomenon

Helitrons are thought to transpose by a semi-replicative rolling-circle (RC) mechanism (Kapitonov and Jurka 2001; Kapitonov and Jurka 2007; Thomas and Pritham 2015) as they encode a protein similar to the replication initiator proteins (Rep) found in RC replicons (Mendiola et al. 1994; Kapitonov and Jurka 2001). The fact that several contigs in *D. virilis* show two or more *DINE-TR1* insertions in tandem arrays (Dias et al. 2015) could be related to this mode of transposition.

Helitron TIs were also identified in organisms such as *Myotis lucifugus* (Pritham and Feschotte 2007), *Daphnia pulex* (Schaack et al. 2010), *Bombyx mori* (Thomas et al. 2010) and *Zea mays* (Du et al. 2008; Yang & Bennetzen 2009). These head-to-tail junctions were explained as a feature of the rolling-circle (RC) transposition mechanism, which is thought to form tandem arrays if the termination signal is bypassed (Thomas et al. 2014). This hypothesis stems from the work conducted by Mendiola et al. (1994) describing the formation of TIs (one-ended transposition) composed of pSU2572 plasmids containing IS91 elements with inactive termination signals. In this case, insertions of whole plasmid units were found fused in a head-to-tail fashion (Fig. 2). This is an expected outcome for the RC transposition mechanism in the case of termination signal bypassing, but only if the donor element is a circular dsDNA. Therefore, an inactive termination signal would only generate TIs of RC transposons (e.g. IS91, *Helitrons*) if the donor element is converted into a 'head-to-tail'

circular unit (Fig. 3B). However, if during transposition the donor TE is located in a chromosome or plasmid, one-ended transposition is expected to capture the 3' flanking sequence, until an alternative termination signal is encountered downstream (Fig. 3A; Feschotte and Wessler 2001).

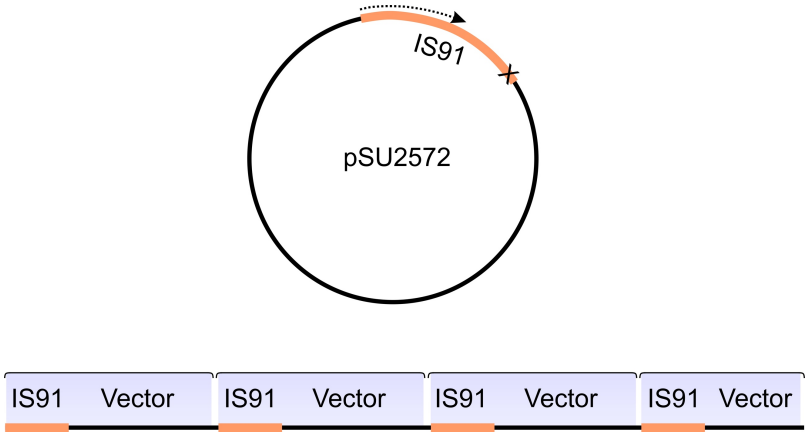


Figure 2. Schematic representation of the tandem insertions observed by Mendiola et al. (1994). When the termination signal is missing at IS91 terminus, the RC transposition loops-out the entire plasmid and generate TIs of the whole construct (IS91 + pSU2572).

The current model used to explain the transposition of *Helitrons* (Fig. 3A; Feschotte and Wessler 2001; Kapitonov and Jurka 2007) is adapted from the first proposed mechanism for RC transposition, using the prokaryotic element IS91 as a model (Mendiola et al. 1994). After the detection of circular ss- and dsDNA intermediates from IS91 (Garcillán-Barcia et al. 2001) a second model was suggested by Garcillán-Barcia et al. (2002) (Fig 3B). They called the first model "concerted" and the second "sequential" (Fig. 3). As noted by Thomas and Pritham (2015), because there is no description of *Helitron* circular DNA species to date, the concerted model has been chosen to explain the transposition mechanism of *Helitrons*, although indirect evidence points to the occurrence of a sequential mechanism in this TE, at least occasionally. For example, the presence of *Helitron* insertions in the form of tandem arrays of the same element (e.g. Pritham and Feschotte 2007; Dias et al. 2015) agrees with the sequential model.

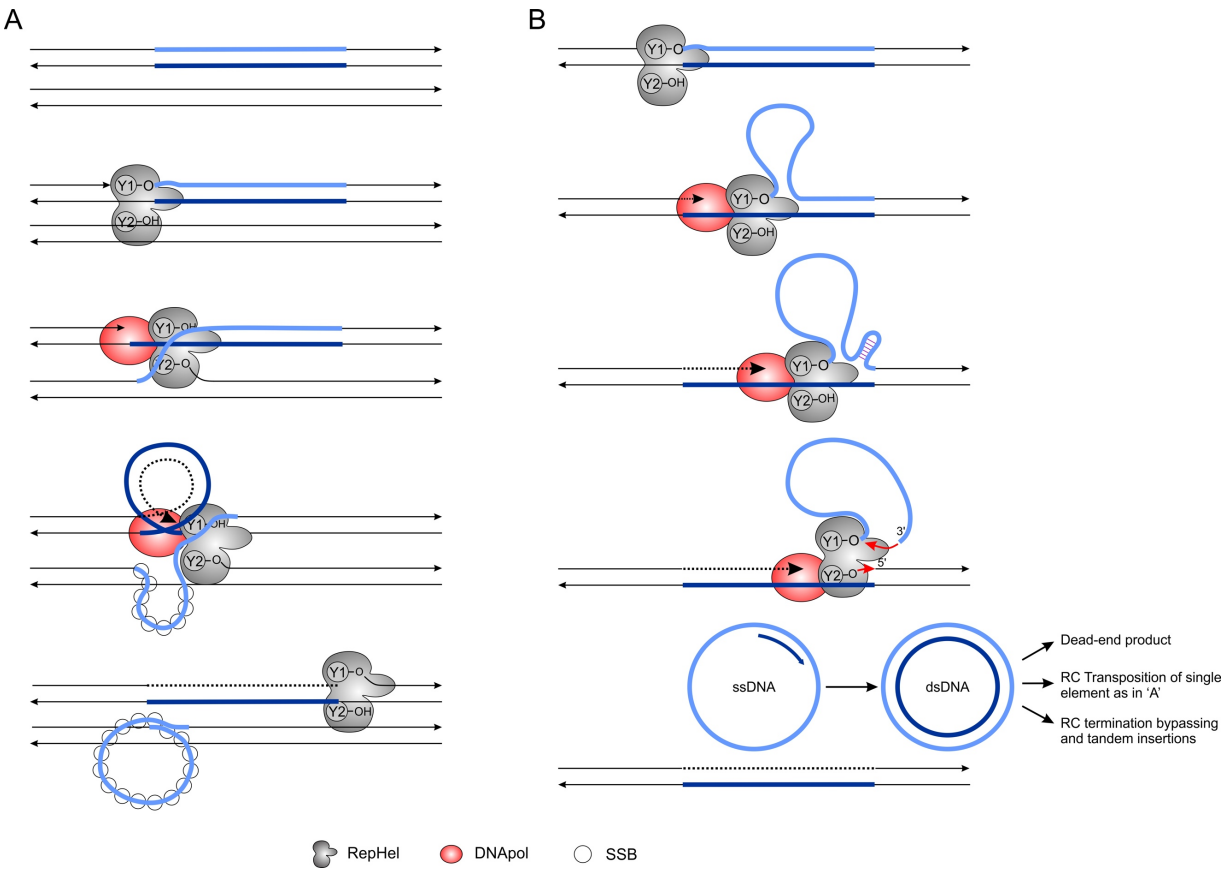


Figure 3. Two RC transposition models proposed for *Helitrons*. (A) The "Concerted" model of RC transposition was proposed for the IS91 prokaryotic elements and evoked to explain *Helitron* transposition (Kapitonov and Jurka 2001; Feschotte and Wessler 2001). Redrawn from Garcillán-Barcia et al (2002). (B) The "Sequential" model of RC transposition was proposed to explain the formation of episomal circular intermediates of IS91. Schematic representation based on the model described in Garcillán-Barcia et al. (2002). Single-stranded binding Proteins (SSBs) were not represented in the B section to improve clarity. RepHel: Replication Initiator Protein and Helicase; DNA Pol: host DNA polymerase.

It is not clear how other typical *Helitron* features, like the tendency to form proximal (but not tandem) insertions or clusters in variable types of arrangements (Du et al. 2008; Yang & Bennetzen 2009; Dias et al. 2015) could be explained by either model alone. Surely, the development of transposition assays will be essential to elucidate all the aforementioned issues; additionally, some important aspects of *Helitron* insertions found in the genomic data have not yet been addressed by the existing models and should be properly outlined in the future.

Perspectives

Given their abundance and wide phylogenetic distribution, *Helitrons* are expected to interact in many ways with their host genomes, and the evolutionary routes and outcomes of such interactions have only recently begun to be unraveled in a few organisms (e.g. Thomas et al. 2010; Ellison and Bachtrog 2013; Barberán-Soler et al. 2014; Carareto et al. 2014; Castanera et al. 2014; Hoffman et al. 2015).

Here we discussed two features displayed by *Helitrons*: the recurrent phenomenon of TIs of entire elements, which seems to be specific to this group of TEs and related to their mobilization mechanism; and the more general evolutionary impact of TR expansion from piRNA-targeted TEs. Advances in the computational identification of *Helitrons* and the development of transposition assays will greatly improve our knowledge about these prolific elements. Notwithstanding, the combined molecular, cytogenetic and computational characterization of *Helitrons* in unexplored genomes will surely keep providing many interesting insights on *Helitron* biology and to the overall impact of repetitive DNA to genome evolution.

Acknowledgements

This work was supported by “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG) (Proc: APQ-01563-14), “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), a doctoral fellowship from “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES) to GBD, and a student fellowship to PH from “Programa Institucional de Auxílio à Pesquisa de Doutores Recém-Contratados da Universidade Federal de Minas Gerais”.

References

Barberán-Soler S, Fontrodona L, Ribó A, Lamm AT, Iannone C, Cerón J, Lehner B, Valcárcel J. Co-option of the piRNA Pathway for Germline-Specific Alternative Splicing of *C. elegans* TOR. Cell Rep 2014; 8(6):1609-16; http://dx.doi.org/10.1016/j.celrep.2014.08.016

Berloco M, Palumbo G, Piacentini L, Pimpinelli S, Fanti L. Position effect variegation and viability are both sensitive to dosage of constitutive heterochromatin in *Drosophila*. G3 (Bethesda) 2014; 4(9):1709-16; PMID: 25053704; http://dx.doi.org/10.1534/g3.114.013045

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 2007; 128:1089-103; PMID: 17346786; http://dx.doi.org/10.1016/j.cell.2007.01.043

Carareto CM, Hernandez EH, Vieira C. Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling *Drosophila* species. Gene 2014; 537(1):93-99; PMID: 24361809; http://dx.doi.org/10.1016/j.gene.2013.11.080

Castanera R, Pérez G, López L, Sancho R, Santoyo F, Alfaro M, Galbadón T, Pisabarro, G, Oquiza JA, Ramírez L. Highly expressed captured genes and cross-kingdom domains present in *Helitrons* create novel diversity in *Pleurotus ostreatus* and other fungi. BMC Genomics 2014; 15(1):1071; PMID: 25480150; http://dx.doi.org/10.1186/1471-2164-15-1071

Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 1994; 371:215-20; PMID: 8078581; http://dx.doi.org/ 10.1038/371215a0

Dias GB, Heringer P, Svartman M, Kuhn GC. *Helitrons* shaping the genomic architecture of *Drosophila*: enrichment of *DINE-TR1* in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosome Res 2015; 23(3):597-613; PMID: 26408292; http://dx.doi.org/10.1007/s10577-015-9480-x

Dimitri P, Pisano C. Position effect variegation in *Drosophila* melanogaster: relationship between suppression effect and the amount of Y chromosome. Genetics 1989; 122(4):793-800; PMID: 2503420

Du C, Caronna J, He L, Dooner HK. Computational prediction and molecular confirmation of *Helitron* transposons in the maize genome. BMC Genomics 2008; 9(1):51; PMID: 18226261; http://dx.doi.org/10.1186/1471-2164-9-51

Elliott TA, Linquist S, Gregory TR. Conceptual and empirical challenges of ascribing functions to transposable elements. Am Nat 2014; 184(1):14-24; PMID: 24921597; http://dx.doi.org/10.1086/676588

Ellison CE, Bachtrog D. Dosage compensation via transposable element mediated rewiring of a regulatory network. Science 2013; 342(6160):846-50; PMID: 24233721; http://dx.doi.org/10.1126/science.1239552

Ferree PM, Barbash DA. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. PLoS Biol 2009; 7(10):2310; PMID: 19859525; http://dx.doi.org/10.1371/journal.pbio.1000234

Feschotte C, Wessler SR. Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci USA 2001; 98:8923-8924; PMID: 11481459; http://dx.doi.org/10.1073/pnas.171326198

Garcillán-Barcia MP, Bernales I, Mendiola M, De La Cruz F. Single-stranded DNA intermediates in IS91 rolling-circle transposition. Mol Microbiol 2001; 39(2):494-502; PMID: 11136468; http://dx.doi.org/10.1046/j.1365-2958.2001.02261.x

Garcillán-Barcia MP, Bernales I, Mendiola MV, de La Cruz F. 2002. IS91 Rolling-circle transposition, p 891-904. In Craig NL, Craigie R, Gellert M, Lambowitz A (ed), Mobile DNA II. ASM Press, Washington D. C. http://dx.doi.org/ 10.1128/9781555817954.ch37

Gregory TR. Synergy between sequence and size in large-scale genomics. Nature Reviews Genetics 2005; 6(9):699-708; PMID: 16151375; http://dx.doi.org/10.1038/nrg1674

Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 2001; 293(5532):1098-102; PMID: 11498581; http://dx.doi.org/10.1126/science.1062939

Hoffmann FG, McGuire LP, Counterman BA, Ray DA. Transposable elements and small RNAs: Genomic fuel for species diversity. Mob Genet Elements 2015; 5(5):1-4; http://dx.doi.org/10.1080/2159256X.2015.1066919

Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 2001; 98(15):8714-8719; PMCID: PMC37501; http://dx.doi.org/10.1073/pnas.151269298

Kapitonov VV, Jurka J. *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trends in Genetics 2007; 23(10):521-529; PMID: 17850916; http://dx.doi.org/10.1016/j.tig.2007.08.004

Kidwell MG, Lisch DR. Transposable elements and host genome evolution. Trends in Ecology & Evolution 2000, 15(3):95-99; PMID: 10675923; http://dx.doi.org/10.1016/S0169-5347(99)01817-0

King DG, Soller M, Kashi Y. Evolutionary tuning knobs. Endeavour 1997; 21(1):36-40; http://dx.doi.org/10.1016/S0160-9327(97)01005-3

Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Research 2008; 16(2):307-324; PMID: 18266060; http://dx.doi.org/10.1007/s10577-007-1195-1

Kuhn GC, Heslop-Harrison JS. Characterization and genomic organization of PERI, a repetitive DNA in the *Drosophila buzzatii* cluster related to DINE-1 transposable elements and highly abundant in the sex chromosomes. Cytogenet Genome Res 2011; 132:79-88; PMID: 20938165; http://dx.doi.org/10.1159/000320921

Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes & development 2013; 27(4):390-399; PMID: 23392610; http://dx.doi.org/10.1101/gad.209841.112

Lee YCG. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. PLoS Genet 2015; 11(6):e1005269; PMID: 26042931; http://dx.doi.org/10.1371/journal.pgen.1005269

Macas J, Koblížková A, Navrátilová A, Neumann P. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 2009; 448(2):198-206; PMDI: 19563868; http://dx.doi.org/10.1016/j.gene.2009.06.014

Martínez-Izquierdo JA, García-Martínez J, Vicient CM. What makes Grande1 retrotransposon different? Genetica 1997; 100(1):15-28; PMID: 9440255; http://dx.doi.org/10.1023/A:1018332218319

Marzo M, Liu D, Ruiz A, Chalmers R. Identification of multiple binding sites for the THAP domain of the Galileo transposase in the long terminal inverted-repeats. Gene 2013; 525(1):84-91; PMID: 23648487; http://dx.doi.org/10.1016/j.gene.2013.04.050

Mendiola MV, Bernales I, De La Cruz F. Differential roles of the transposon termini in IS91 transposition. Proc Natl Acad Sci USA 1994; 91(5):1922-26; PMID: 8127907; http://dx.doi.org/10.1073/pnas.91.5.1922

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Res 2015; 23(3):583-96; PMID: 26293606; http://dx.doi.org/10.1007/s10577-015-9483-7

Morales-Hojas R, Reis M, Vieira CP, Vieira J. Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Mol Phylogenet Evol 2011; 60(2):249-58; PMID: 21571080; http://dx.doi.org/10.1016/j.ympev.2011.04.022

Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. Chromosoma 2014; 123(4):313-25; PMID: 24763964; http://dx.doi.org/10.1007/s00412-014-0462-0

Pritham EJ, Feschotte C. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci USA 2007; 104(6):1895-900; PMID: 17261799; http://dx.doi.org/10.1073/pnas.0609601104

Šatović E, Plohl M. Tandem repeat-containing MITEs in the clam *Donax trunculus*. Genome Biol Evol 2013; 5(12):2549-59; PMID: 24317975; http://dx.doi.org/10.1093/gbe/evt202

Scalvenzi T, Pollet N. Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA. Mol Phylogenet Evol 2014; 81:1-9; PMID: 25193611; http://dx.doi.org/10.1016/j.ympev.2014.08.014

Schaack S, Choi E, Lynch M, Pritham EJ. DNA transposons and the role of recombination in mutation accumulation in *Daphnia pulex*. Genome Biol 2010; 11(4):R46; PMID: 20433697; http://dx.doi.org/10.1186/gb-2010-11-4-r46

Shapiro JA, von Sternberg R. Why repetitive DNA is essential to genome function. Biological Reviews 2005; 80(02):227-250; PMID: 15921050; http://dx.doi.org/10.1017/S1464793104006657

Thomas J, Pritham E. *Helitrons*, the Eukaryotic Rolling-circle Transposable Elements. Microbiol Spectr 2015; 3(4):MDNA3-0049-2014; PMID: 26350323; http://dx.doi.org/10.1128/microbiolspec.MDNA3-0049-2014

Thomas J, Schaack S, Pritham EJ. Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol Evol 2010; 2:656-64; PMID: 20693155; http://dx.doi.org/10.1093/gbe/evq050

Thomas J, Vadnagara K, Pritham EJ. *DINE-1*, the highest copy number repeats in *Drosophila melanogaster* are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). Mob DNA 2014; 5(1):18; PMID: 24959209; http://dx.doi.org/10.1186/1759-8753-5-18

Yang HP, Barbash DA. Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. Genome Biol 2008; 9(2):R39; PMID: 18291035; http://dx.doi.org/10.1186/gb-2008-9-2-r39

Yang L, Bennetzen JL. Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. Proc Natl Acad Sci USA 2009; 106:19922-27; PMID: 19926865; http://dx.doi.org/10.1073/pnas.0908008106

Zayed H, Izsvák Z, Walisko O, Ivics Z. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. Mol Ther 2004; 9(2):292-304; PMID: 14759813; http://dx.doi.org/10.1016/j.ymthe.2003.11.024

Zhang H, Koblížková A, Wang K, Gong Z, Oliveira L, Torres GA, Wu Y, Zhang W, Novák P, Buell CR, Macas J, Jiang, J. Boom-bust turnovers of megabase-sized centromeric DNA in Solanum species: rapid evolution of DNA sequences associated with centromeres. Plant Cell 2014; 26(4):1436-47; PMID: 24728646; http://dx.doi.org/10.1105/tpc.114.123877

Zuckerkandl E. A possible role of "inert" heterochromatin in cell differentiation. Action of and competition for "locking" molecules. Biochimie 1974; 56:937-54; PMID: 4614863

Capítulo 3

Improved assembly and characterization of 172TR, a family of euchromatic tandem repeats in the euchromatin of *Drosophila virilis*

Este capítulo é composto de um manuscrito ainda não submetido para publicação. Nele, realizamos a caracterização de uma família de repetições em tandem identificadas em 2013 no genoma sequenciado de *D. virilis*. Utilizamos citogenética molecular e dados genômicos de Sanger e Illumina para caracterizar estas repetições. Além disso, produzimos uma nova montagem do genoma de *D. virilis* utilizando dados de sequenciamento de terceira geração (PacBio). Comparamos a eficiência desta nova montagem em relação ao genoma de referência na representação das cadeias dessa família repetitiva. Parte dos dados contidos neste manuscrito foram obtidos entre abril e julho de 2017, em estágio realizado na Universidade da Geórgia sob supervisão do Dr. Casey Bergman, e fazem parte de um projeto em andamento.

Improved assembly and characterization of 172TR, a family of euchromatic tandem repeats in the euchromatin of *Drosophila virilis*

Guilherme B. Dias[1], Bráulio Leão[1], Marta Svartman[1], Justin P. Blumenstiel[2], Casey M. Bergman[3], Gustavo C. S. Kuhn[1]

1 - Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
2 - Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA
3 - Department of Genetics and Institute of Bioinformatics, University of Georgia, Athens, GA, USA

Keywords: tandem repeats, minisatellite, *Drosophila virilis*, PacBio, genome assembly

Abstract

Genomic regions containing highly similar tandem repeats (TRs) whose monomers approach or surpass read length are particularly hard to assemble. As a result, current genome assemblies are highly fragmented at these regions, and the analysis of TR array structure and evolution is hindered. Herein, we allied molecular cytogenetics, first and second-generation genome sequencing datasets, and the ultra-long read sizes of Single Molecule Real-Time (SMRT) sequencing to characterize an abundant family of TRs from *Drosophila virilis*. The consensus sequence of this repeat family is 172 bp long, and they are henceforth referred to as 172TR. Our results indicate that the 172TR family is an unusually large and abundant (~15500 copies) minisatellite from *D. virilis*. This repeat is shared between species of the *virilis* subgroup, and its genomic abundance varies 6-fold amongst all species. These repeats are exclusively euchromatic and enriched in the distal regions of the chromosomes in *D. virilis*. Small RNAs derived from the 172TRs are abundant in embryos and gonads. SMRT sequencing proved to be considerably more efficient than the Sanger reference genome in recovering fully resolved arrays of the 172TRs.

Introduction

Amongst the repeats that populate eukaryote genomes, tandem repeats (TRs) are major contributors (Melters et al. 2013). They can exist either in very large arrays concentrated in the heterochromatic regions of chromosomes, where they are called satellite DNAs, or in smaller arrays widespread along the chromosome arms, where they are called micro or minisatellites (Richard et al. 2008). Although the large blocks of centromeric and subtelomeric satellite DNAs are usually not captured in genome assemblies, they are not the only TRs that generate sequencing gaps - euchromatic TRs are widespread and can also contribute to the current fragmented state of genome assemblies (Treangen and Salzberg 2012; Baker 2012). Both Sanger sequencing reads (~800 bp) and next generation sequencing reads (~100 bp) proved to be insufficient to bridge large and highly similar repetitive regions. These suboptimal genome assemblies preclude researchers from gaining insight into the large-scale structure of genome sequence, as well as the local organization of repeat arrays.

New approaches in sequencing are offering promising ways of obtaining improved genome assemblies. These so called third generation sequencing technologies capitalize on ultra-long reads, frequently larger than 10 kb, to bridge repeats and produce highly contiguous assemblies (Schadt et al. 2010). The most established of these technologies is the Single Molecule Real Time (SMRT) sequencing, developed and commercialized by Pacific Biosciences (PacBio, Menlo Park, USA). SMRT sequencing involves the use of a proprietary DNA polymerase and monitoring fluorescence released after nucleotide incorporation. The current SMRT sequencing chemistry and hardware is able to produce runs where half of the data is present in reads larger than 20 kb, and the largest reads may surpass 60 kb (http://www.pacb.com/smrt-science/smrt-sequencing/).

The genus *Drosophila* has long been in the forefront of genomics (Ashburner and Bergman 2005), and there are currently 22 whole-genome assemblies for species of this group available in FlyBase (http://flybase.org). Although most *Drosophila* species sequenced so far have modest genome sizes (~170 Mbp) and repeat contents (5-25%), *Drosophila virilis* stands out as having one of the largest genomes (~ 360 Mbp), and largest heterochromatin and repeat proportions in the genus (50%; *Drosophila* 12 Genomes Consortium 2007; Mahan and Beck 1986). Several families of both dispersed and tandem repeats were identified in the genome of *D. virilis* so far (Gall et al. 1974; Feschotte et al. 2009; Abdurashitov et al. 2013; Dias et al. 2014, 2015). However, few of these works have performed in depth characterizations of these repeat families, partly, because of the incomplete nature of the reference genome regarding repetitive sequences.

Herein, we aimed to better characterize a TR family identified in the genome sequence of *Drosophila virilis* (Abdurashitov et al. 2013). The consensus sequence for this family has 172 bp, and it is henceforth referred to as 172TR. We combined molecular cytogenetics and existing genomic resources for *D. virilis*, including a Sanger reference genome (RG) and short read sequencing to characterize these repeats. In addition, we generated a de novo genome assembly with SMRT sequencing data. Our results indicate that the 172TR family is an unusually large and abundant minisatellite sequence present in the *virilis* species subgroup. These repeats are euchromatic and enriched in the distal regions of the chromosomes, and there is a considerable enrichment of small RNAs derived from them in the gonadal tissues of *D. virilis*. The PacBio assembly (PBA) proved to be considerably more efficient than the RG in recovering fully resolved arrays of 172TRs. This new genomic resource will aid in answering many biological questions, as well as allowing the validation of TR array assembly.

Material and methods

Genome sequencing and assembly

Flies from the *D. virilis* strain 160 were inbred by sibling mating for ten generations. Genomic DNA was extracted from 12h starved adult female flies with the Qiagen Blood Cell and Culture midi kit according to the manufacturer's instructions (Qiagen, Hilden, Germany). DNA was then sent to the University of Michigan sequencing core where it was sheared to 10-20 kb fragments with the Covaris g-TUBEs system (Covaris, Woburn, USA). DNA was sequenced using 21 SMRT cells with P6-C4 chemistry on a PacBio RS II instrument.

A total of 3,700,363 raw subreads were produced, summing ~29 billion base pairs and representing 82x coverage. This dataset was used for de novo genome assembly with the Canu software, version 1.5, with default parameters (Koren et al. 2017). The genomeSize parameter was set to 200 Mb. This parameter controls the amount of reads to be corrected, summing 40x coverage from the longest to the shortest reads. Increasing values for the genomeSize parameter quadratically increased the computational time and did not result in better assembly statistics with our read dataset.

Assembly quality and contiguity were evaluated by using the N50/L50 statistics together with whole genome alignments against the reference *D. virilis* genome. The whole genome alignment was performed and plotted with the MUMmer package (Delcher et al. 2002). The parameters for the alignment were: min match -l 100; and min cluster -c 1000.

Chromosome mapping of 172TRs

Salivary glands and neuroblasts of third instar larvae from *D. virilis* (strain 15010-1051.51) were dissected to obtain polytene and metaphase chromosome preparations, respectively (Ashburner 1989; Baimai 1977). The probes used for FISH were obtained by PCR with a set of specific primers (forward ATTTATGGGCTGGGAAGCTTTGACGTATG, and reverse CGGTCAAATCTCATCCGATTTTCATGAGG), and labeled by nick translation with the DIG-nick translation mix (Roche Applied Science, Penzberg, Germany). Hybridizations were performed as described in Dias et al. (2014).

Mapping of small RNAs to the 172TR

We utilized a small RNA sequencing datasets generated by Rozhkov et al. (2010; Bioproject: PRJNA127489) to perform read mapping against the consensus of the 172TR

family (Abdurashitov et al. 2013). These datasets included small RNAs from early embryos (0-2h), gonads and adult carcasses (without the gonads). Mapping was performed with Lastz (Harris 2007), implemented on the Galaxy platform (http://usegalaxy.org; Giardine et al. 2005; Goecks et al. 2010), and excluded reads with less than 85% similarity to the 172TR consensus.

172TR genome annotation

The *D. virilis* reference genome (strain 15010-1051.87) and associated information was obtained from the University of California at Santa Cruz website (UCSC; http://genome.ucsc.edu/cgi-bin/hgGateway?db=droVir2). We utilized the version 2 (droVir2) because this is the latest version supported by the UCSC Genome Browser.

The 172TRs were annotated in both the RG and the PBA using the RepeatMasker software, version 4.0.5 (Tarailo-Graovac and Chen 2009). The consensus sequence of the 172TR family published in Abdurashitov et al. (2013) was used as the custom repeat library. The search engine used was the wu-blast in sensitive mode (http://blast.wustl.edu). Because RepeatMasker utilizes a consensus sequence to annotate similar repeats in the genome, stretches lacking similarity between the consensus and the target sequence often produce fragmented alignments and inflate estimates of array number. To remove this artifact, we joined nearby alignments (<=100 bp away, in the same orientation) using the BEDTools software (Figure 1; Quinlan and Hall 2012).

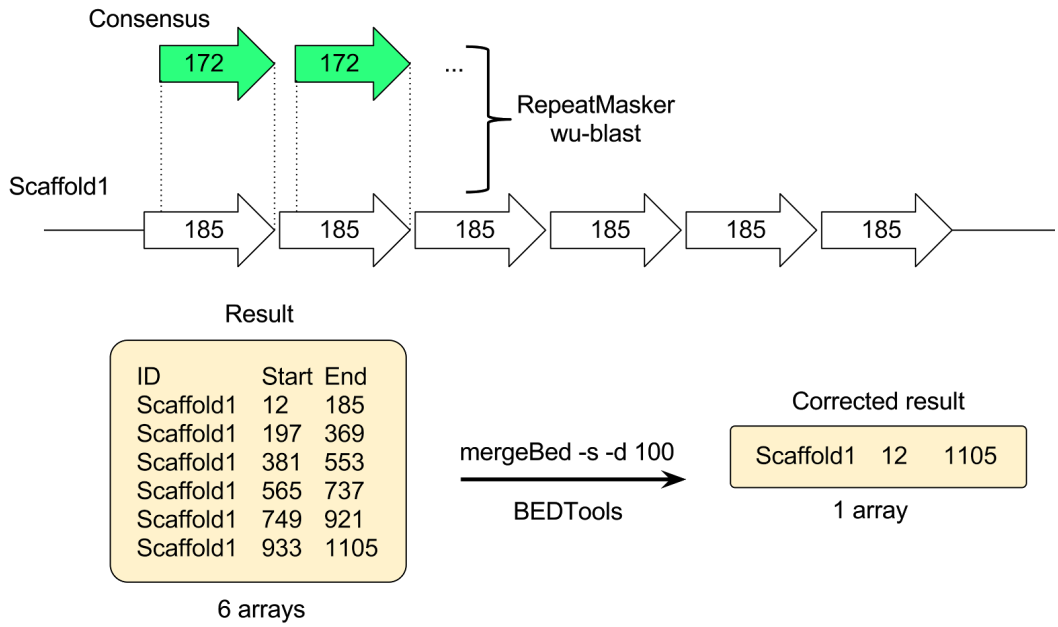


Figure 1. Pipeline for joining nearby hits of the 172TRs using RepeatMasker and BEDTools.

We analyzed the size distribution of all arrays in the RG and the PBA using BEDTools and BEDOPS (Figure 2; Neph et al. 2012). The results were plotted using RStudio (https://www.rstudio.com/products/rstudio/download/) and Inkscape (https://inkscape.org/pt-br/).

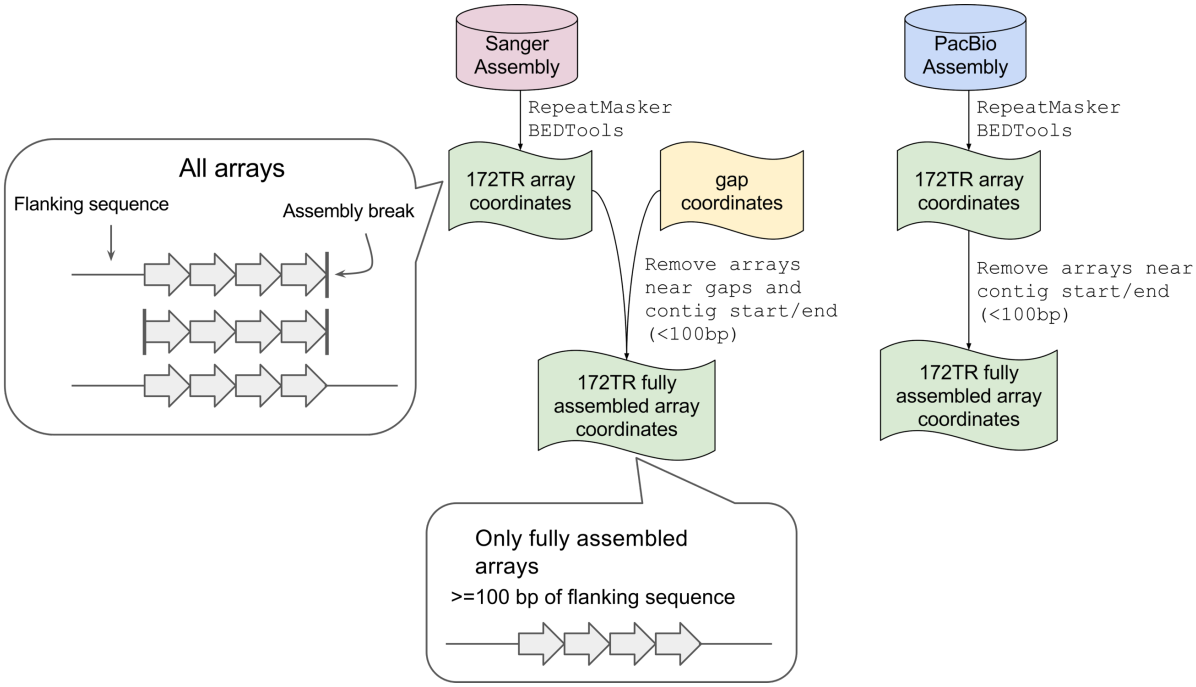


Figure 2. Annotation pipeline for the 172TRs in the reference genome and the PacBio assembly.

172TR abundance estimation

We estimated 172TR genomic abundance in several datasets from *D. virilis* and other species of the *virilis* subgroup. We used the RepeatMasker software and the 172TR published consensus to analyze both random samples of sequencing reads and genome assemblies. RepeatMasker was run with wu-blast and the sensitive setting. Random samples of reads were obtained with the seqtk toolkit (https://github.com/lh3/seqtk). The Sanger reads from *D. virilis* were downloaded from the ftp website of the NCBI Trace Archive. Accession numbers for the short reads are: *D. americana*, ERR957820, ERR957822, and SRR5278983; *D. lummei*, SRR5278982; *D. novamexicana*, SRR5278981; *D. virilis*, SRR5278980, SRR1536176, SRR1200631, SRR1536175, and SRR1200817.

Results and discussion

172TRs are exclusively euchromatic, highly dispersed, and enriched in the distal regions

To obtain an assembly-independent overview of 172TR chromosome distribution, we performed FISH with 172TR probes to the polytene and metaphase chromosomes of *D. virilis*. The results indicate an exclusive euchromatic localization of 172TRs, present in all chromosomes except the Y and the dot (Figure 3). This analysis also evidenced an enrichment of 172TRs on the distal regions of the chromosome arms.

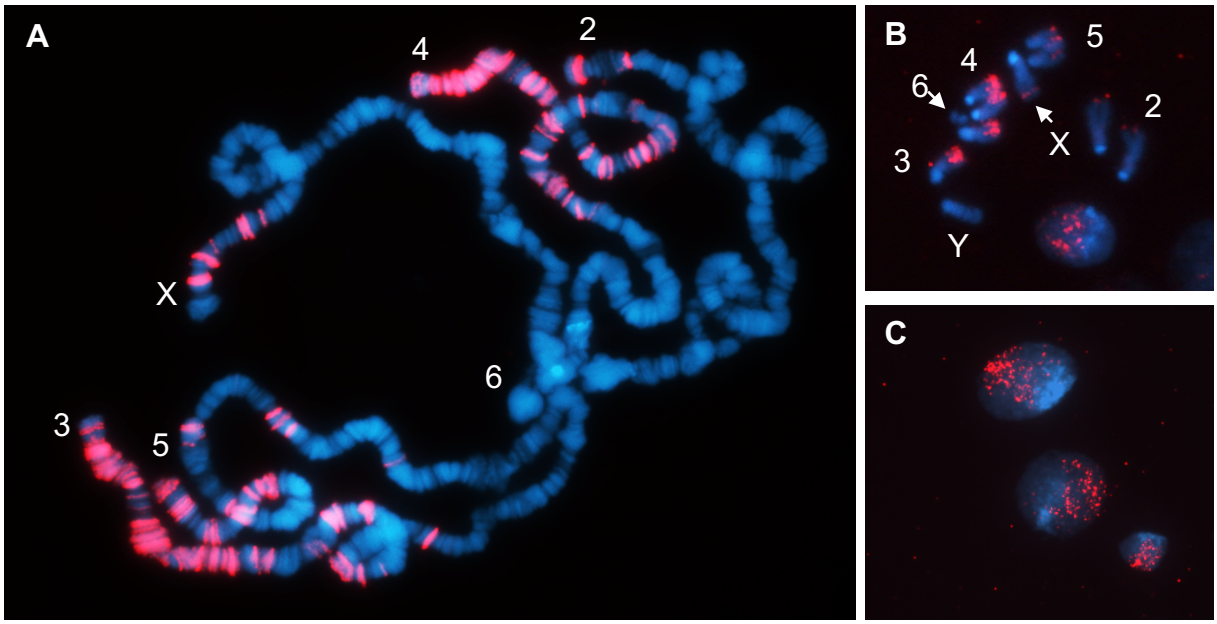


Figure 3. FISH of 172TR probes onto the polytene (A) and metaphase (B) chromosomes, and interphase nuclei (C) of *D. virilis*. Polytene chromosome arms were identified based on the landmark regions defined in Gubenko and Evgen'ev (1984).

A somewhat similar distribution of euchromatic TRs was also observed in species of the *Drosophila ananassae* subgroup, where a similar sized TR (175-200 bp) is also spread along the chromosome arms (Nozawa et al. 2006). The monomer size and chromosome distribution of these TRs are not compatible with the definitions of either microsatellites or satellite DNAs. Based on the definition proposed by Tautz (1993), the most suitable classification for the 172TRs is that of a minisatellite. However, 172TR monomers are considerably larger than the usual minisatellite range (10-100 bp). Also, its arrays may reach more than 42 kb (see below), substantially larger than the 30 kb upper length described for

minisatellite arrays (Armour and Jeffreys 1992). Although the features of the 172TR family seem to be somewhat atypical for a minisatellite, it is important to acknowledge that the minisatellites literature is heavily biased by the use of these sequences as individual markers (Jeffreys 1987).

Although their distribution clearly indicates that these TR arrays are highly mobile, there is no current consensus on how they could jump. An interesting hypothesis is that monomers within an array may suffer intrastrand homologous recombination, forming extrachromosomal circular DNAs (eccDNAs; Walsh 1987). These molecules could, in turn, reintegrate elsewhere in the genome via illegitimate recombination. Although not necessary, an additional step involving rolling circle replication of TR eccDNAs could also take place (Cohen and Seagal 2009). Although eccDNAs derived from TRs were found in every species analyzed to date, there is no evidence yet for the reintegration of these circles in the genome other than the widespread distribution of TR arrays in euchromatic regions.

172TR-derived small RNAs are abundant in gonads and early embryos of *D. virilis*

Even though most repeats are non-protein-coding, these sequences may often give rise to RNAs which can be processed into one of a few classes of small interfering RNAs (Kim et al. 2009). These small RNAs are associated with several biological roles including transposon silencing, chromatin modulation, DNA repair and gene regulation (reviewed in Castel and Martienssen 2013). To investigate whether 172TRs are transcribed and generate small processed RNAs, we utilized a small RNA dataset available for the strain 160 of *D. virilis* (Rozhkov et al. 2010).

Although it is not feasible to map short sequencing reads to specific loci of repetitive elements because of their high similarity, we can estimate the general abundance of 172TR-derived small RNAs by mapping these RNAseq datasets to a consensus sequence of the 172TR family. This analysis revealed an enrichment of small RNAs matching the 172TRs in early embryos (0-2h) and both testes and ovaries of adult *D. virilis* flies (Figure 4A). The embryonic and gonadal enrichment of 172TR-derived small RNAs could indicate that these molecules belong to the piwi interacting RNAs (piRNAs) class (Brennecke et al. 2007). However, a typical feature of piRNAs is the presence of a strong uridine bias in their 5' end (Aravin et al. 2003; Brennecke 2007). This bias was indeed observed in the testes and ovaries, but not in embryos (Figure 4B). Furthermore, different classes of small RNAs have different size profiles, so we analyzed the size distribution of the 172TR RNAs in embryos and gonads of *D. virilis*. Although embryos, testes, and ovaries present RNAs with fairly similar mean sizes (23, 23.7, and 24.7 nt, respectively), the shape of the size distributions varies

greatly (Figure 4C, D, and E). The RNAs from testes are the most heterogeneous in size, with no accentuated peak (Figure 4B). The embryos and ovaries distributions on the other hand, are clearly peaked at 22 and 27 nt, respectively (Figure 4A, and C).

The wide variation in read sizes of RNAs derived from the 172TRs could indicate that more than one population of small RNAs is present in each of these datasets. In fact, considering the broad distribution of 172TR arrays along *D. virilis* chromosomes, it is not difficult to conceive that different small RNAs may derive from different loci. The only tissue displaying a somewhat clearer small RNA profile are the ovaries. This tissue presents a strong 5' uridine bias and a size distribution peaked at 27 nt, indicating that these molecules may belong to the piRNA class (Brennecke et al. 2007). However, even for this dataset, other small RNA types may be present. The analysis of small RNA datasets collected at several additional time points during development as well as performing depletion of such molecules in cell culture and whole flies will aid in the functional characterization of this TR family.

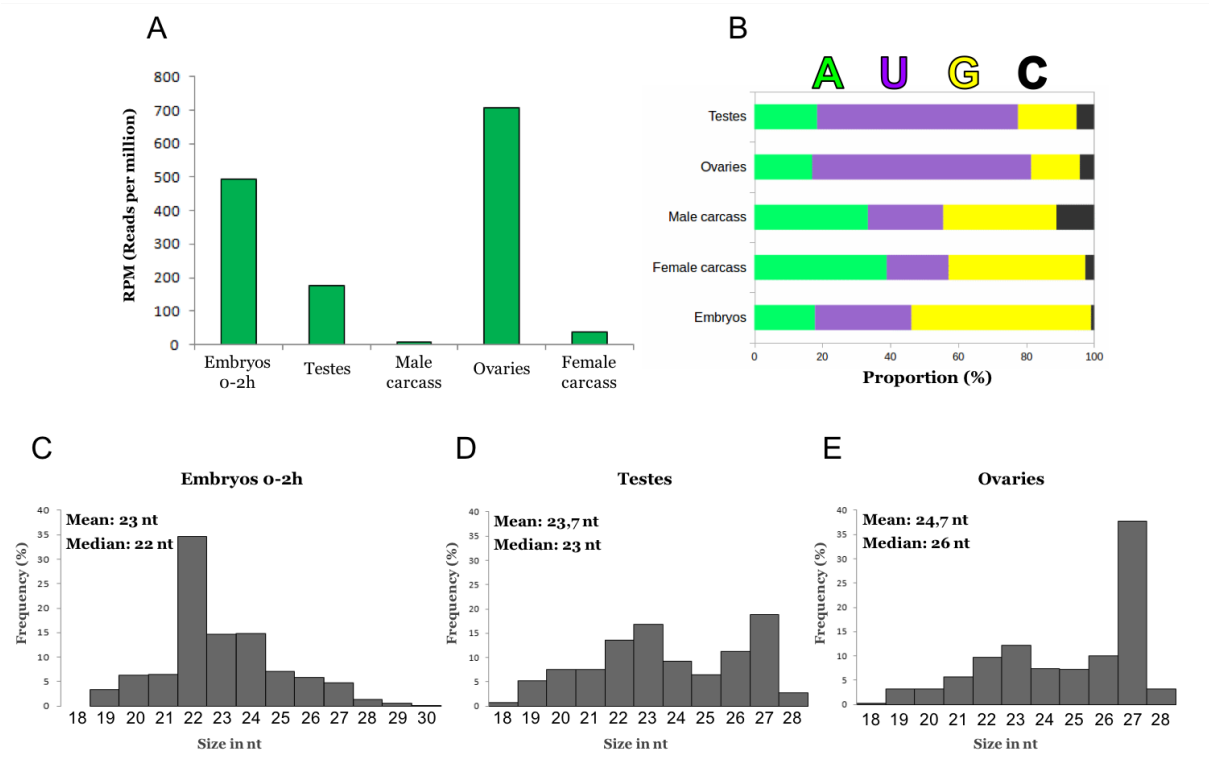


Figure 4. A: Small RNA profile of the 172TRs in tissues from *D. virilis*. RPM, Reads per million. B: 5' composition of the RNAs matching the 172TRs in *D. virilis*. C, D and F: Read size in nucleotides (nt) of the small RNAs matching 172TRs in embryos, testes and ovaries, respectively.

Improving *D. virilis* assembly contiguity with SMRT sequencing

Regions containing highly similar tandem repeats (TRs) result in highly fragmented genome assemblies, and the analysis of TR array structure and evolution is hindered (Baker 2012; Miga 2014). Herein, we took advantage of the ultra-long read size from SMRT sequencing to obtain a more contiguous assembly of the *D. virilis* genome, focusing on the annotation of the 172TRs.

A total of 3,700,363 subreads were generated, amounting to 29.8 Gbp and representing ~82x coverage of the *D. virilis* genome. Read size distribution is given in figure 5. Because SMRT sequencing reads have high error rates, a correction step needs to be performed prior to assembly. We performed self-correction and assembly of PacBio reads without the use of second generation data, using the software Canu version 1.5 (Koren et al. 2017). By specifying 200 Mb as the genomeSize parameter, a total of 533,004 corrected reads were obtained, amounting to ~7 Gbp. After the trimming step, assembly proceeded with 520,866 reads (6.8 Gbp). Assembly took 82 hours (~1300 CPU hours) on a 28-processor node with 256 Gb of RAM. The resulting assembly statistics are given in Table 1.

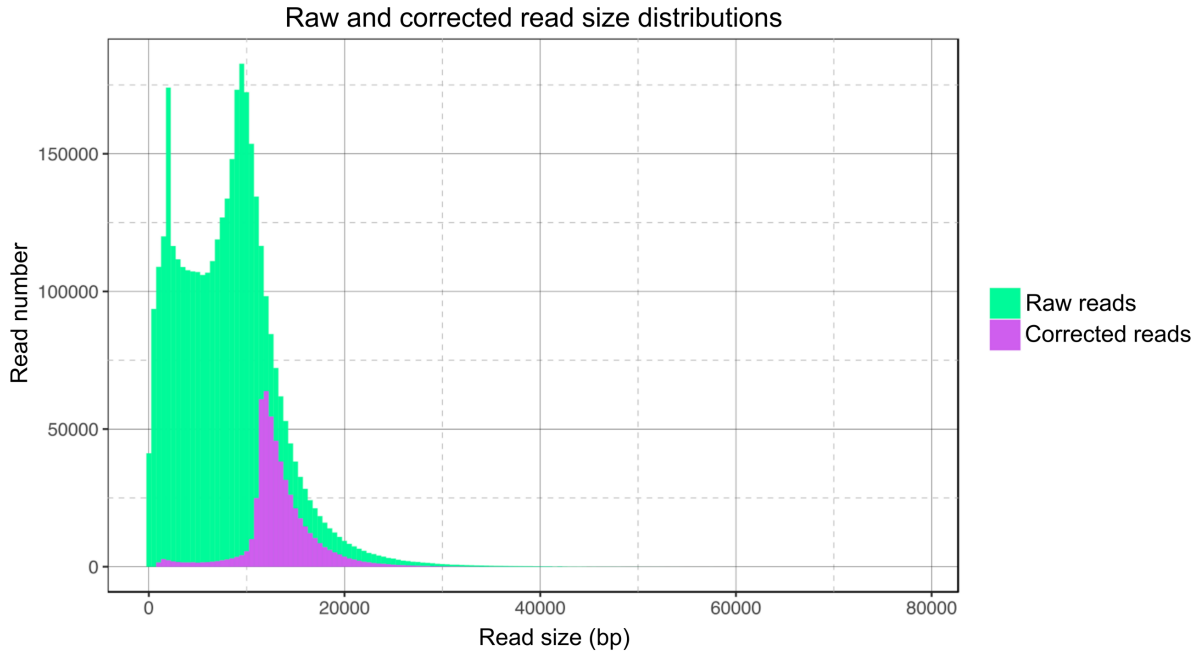


Figure 5. Raw and corrected PacBio read size distribution. Reads were self-corrected with Canu 1.5 and a 200 Mb genomeSize setting.

Table 1. Statistics of the *D. virilis* reference genome and the new PacBio assembly.

	Reference Genome	**PacBio assembly**
Coverage	8x	82x
Assembly size (bp)	206,026,697	169,694,069
Sequencing gaps (bp)	16,820,834	-
Scaffold number	13,530	-
Scaffold N50	10,161,210	-
Scaffold L50	6	-
Contig number	18,382	189
Contig N50	120,091	6,841,436
Contig L50	389	7

The PBA encompassed ~170 Mbp, likely covering most of the euchromatin. This assembly contained 189 contigs, with a contig N50 of 6,8 Mbp. This represents a 56x increase in the contig N50 relative to the RG (120,091 bp). Although these assemblies derive from different strains of *D. virilis*, whole genome alignments between the PBA and the RG evidenced a great overall collinearity (Figure 6). Differences between the RG and PBA could indicate errors in either assembly, or they could reflect true rearrangements of some regions between strains 15010-1051.87 (RG) and 160 (PBA). Because the goal of the PBA was to better annotate the 172TRs, no further validation was performed regarding inconsistencies with the RG.

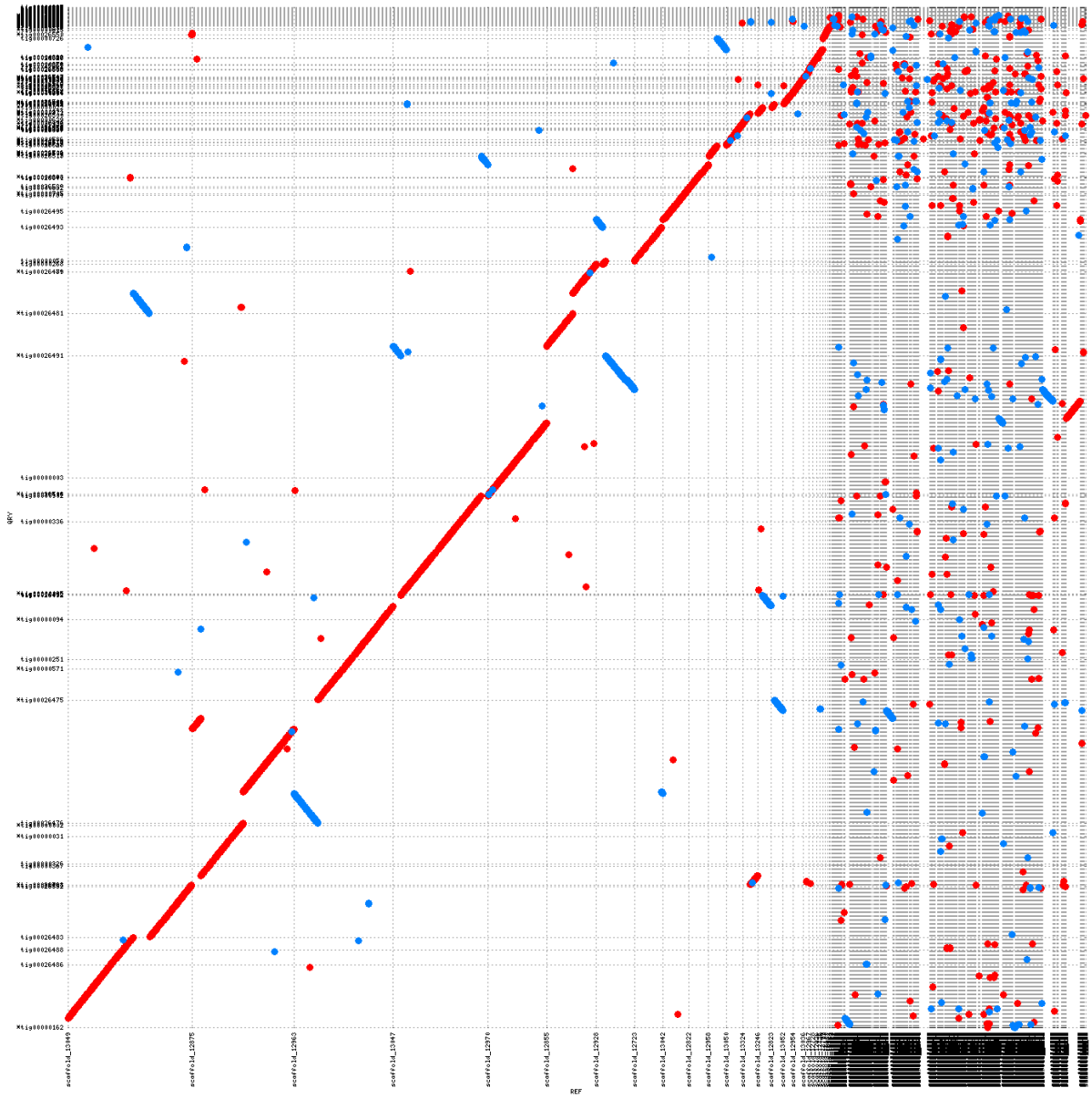


Figure 6. MUMmerplot of the whole genome alignment between the *D. virilis* reference genome (x axis) and the PacBio assembly (y axis). Red diagonals indicate segments in the same orientation in both assemblies, whereas blue diagonals indicate inverted segments.

The use of SMRT sequencing already proved to be efficient in obtaining high quality de novo assemblies for several organisms (e.g. Daccord et al. 2017; Lok et al. 2017; Zimin et al. 2017). In *Drosophila*, only a few species currently have assemblies derived from PacBio data. These include *D. melanogaster*, *D. pseudoobscura* and *D. serrata*. For *D. pseudoobscura*, low coverage PacBio reads were incorporated to the original Sanger assembly with the aim of closing gaps (English et al. 2012). This improved the contig N50 from 53,051 bp to 224,350 bp. For *D. melanogaster*, full de novo PacBio assemblies with ~90x coverage using FALCON and the PBcR-CA pipelines reached contig N50 of 5 Mbp and

15,2 Mbp, respectively (Koren et al. 2012; 2013; Chin et al. 2013; https://github.com/PacificBiosciences/DevNet/wiki/*Drosophila*-sequence-and-assembly). For *D. serrata*, 65x PacBio coverage was assembled with the MHAP algorithm and the Celera Assembler, reaching a contig N50 of 942,627 bp (Berlin et al. 2015; Allen et al. 2017). *D. virilis* PBA reached very good assembly statistics, even when compared to scaffolded Sanger genomes (*Drosophila* 12 Genomes Consortium 2007) or other PacBio assemblies of smaller and less repetitive genomes (e.g. *D. melanogaster* and *D. serrata*). As expected, the PBA performs vastly better than second generation de novo assemblies from Illumina data alone, such as that from *D. americana*. This species belongs to the *virilis* subgroup, has a smaller genome than *D. virilis*, and its contig N50 reached 9,249 bp and 14,233 bp for the strains W11 and H5, respectively (Fonseca et al. 2013). The *D. virilis* PBA performed 480x better than this Illumina-only de novo assembly of a close species.

The total length of the *D. virilis* PBA is considerably smaller than the RG (170 against 206 Mbp). Part of this difference is explained by the high number of sequencing gaps in the RG, which span an estimated 16,8 Mbp (Table 1). The PBA on the other hand, is not organized in scaffolds, and thus contains no sequencing gaps. In addition to that, the RG contains a large amount of very short scaffolds, comprised nearly exclusively of repetitive elements. These scaffolds cannot be confidently mapped to chromosome arms and could represent collapsed repeats; thus, their correctness is hard to ascertain (Treangen and Salzberg 2012). The Canu assembly pipeline places these low confidence repetitive sequences out of the main assembly. In fact, when considering only the scaffolds from the RG that are larger than 200 kb, this assembly size gets much closer to the PBA size (172 Mbp; *Drosophila* 12 Genomes Consortium 2007), and the PBA is considerably less fragmented than the former.

172TR arrays in the reference genome and the PacBio assembly

The new *D. virilis* PBA proved to be of high quality compared to first and second-generation assemblies and also compared to full PacBio de novo assemblies of other *Drosophila* species. With this new resource, we proceeded to annotate the 172TR arrays in both the RG and the PBA. By using RepeatMasker we determined the RG and the PBA to contain a comparable overall amount of sequences similar to the 172TR consensus (1,794,696 bp and 1,812,914 bp, respectively). In the RG, however, these sequences are divided amongst 935 arrays, whereas in the PBA they are divided into only 362 arrays (Table 2). This dataset comprises all arrays, including the ones that are broken either by physical or sequencing gaps (Figure 7).

Table 2. 172TR annotation in the reference genome and the new PacBio assembly.

	Reference Genome	PacBio Assembly
Total abundance (bp) [a]	1,794,696	1,812,914
Total array number [b]	935	362
Abundance in fully assembled arrays (bp) [c]	541,084	1,672,363
Fully assembled arrays number [d]	291	352

[a] Total amount of bp masked by RepeatMasker. [b] Number of arrays of at least 172 bp. [c] Total amount of bp in arrays that are fully assembled, that is, containing at least 100 bp of flanking sequence on each side. [d] Number of arrays of at least 172 bp that are fully assembled.

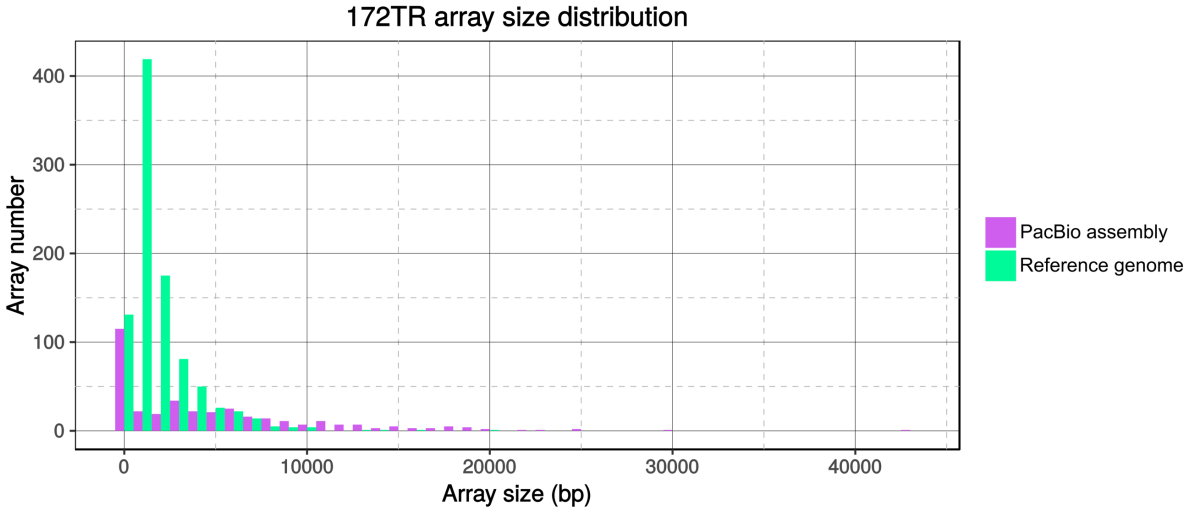


Figure 7. 172TR array size distribution in the PacBio assembly and the reference genome. This analysis includes all arrays with at least 172 bp.

When considering only the fully assembled arrays, that is, the ones with at least 100 bp of flanking sequence on both sides, it becomes clear that the PBA performs strikingly better (Table 2; Figure 8). In fact, 68% of the arrays from the RG are broken by physical or sequencing gaps, preventing us from knowing their real size and structure. The PacBio assembly, on the other hand, has no sequencing gaps and produced a highly contiguous set of arrays, with less than 3% being broken by physical gaps (Table 2). Our improved assembly recovered three times as much data in fully resolved arrays compared to the RG (1.67 Mbp vs 0.54 Mbp), even though the PBA contains only a 20% surplus of complete arrays. This suggests that in addition to fragmented assembly, the RG also has a high level of read collapsing (Table 2). The largest fully assembled array in the PBA spans 42 kb, while the largest one from the RG spans only 10 kb.

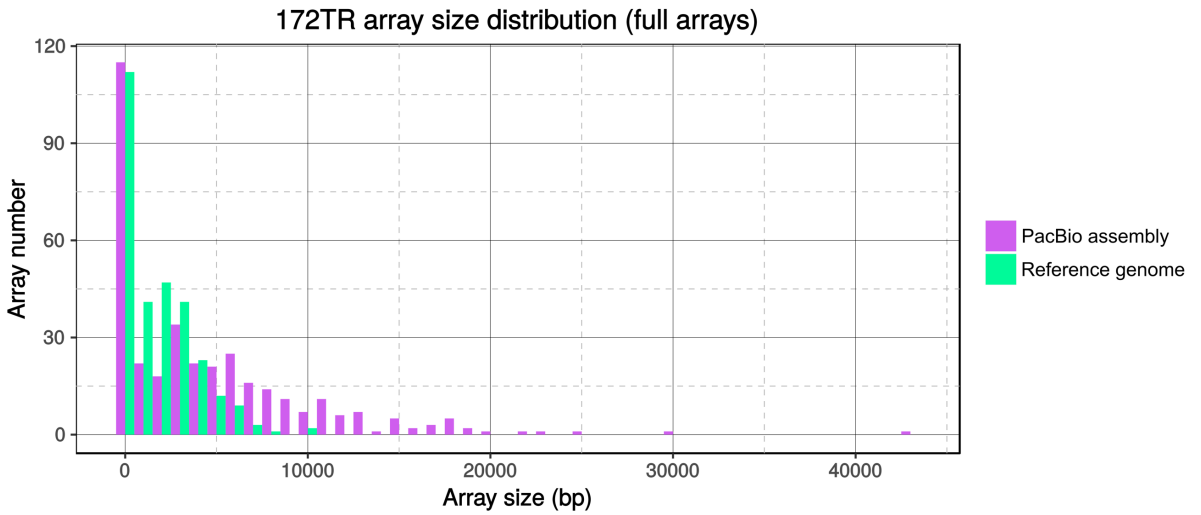


Figure 8. Size distribution of fully assembled 172TR arrays in the PacBio assembly and the reference genome. This analysis includes all arrays at least 172 bp long that possess 100 bp flanking sequence on both sides.

The possibility of read collapsing in the RG is further supported by the analysis of a specific 172TR array residing between the *apterous* and *unc-5* genes. This locus was assembled twice: first, by subcloning and Sanger capillary sequencing of a single P1 clone (Bergman et al. 2002); and second, by Sanger whole-genome shotgun sequencing (*Drosophila* 12 Genomes Consortium). When comparing the array size in these two assemblies to our PBA, both the P1 clone assembly and the PBA present the same array size of ~8 kb (or 46 monomers). The RG, however, present an array of only ~3 kb and 18 monomers. This could indicate the collapsing of similar reads and the underestimation of array size in the RG. The recovery of fully assembled arrays is important as it allows downstream validation analyses of these size inconsistencies with methods such as long-range PCR or fiber-FISH.

172TR abundance in *D. virilis*

To better evaluate how complete our PBA is in its representation of the 172TRs, we estimated the 172TR abundance in several genomic datasets available. This analysis revealed considerable variation in the 172TR abundance estimates, with a 25% coefficient of variation (Table 3). The values found for both assemblies cannot be compared to the sequencing read estimates because the former are known to be incomplete, while the latter are assumed to represent the entire genome. Amongst the read estimates there is a smaller, but still considerable, coefficient of variation (17%). This variation could arise from a number of reasons including the presence of polytene tissues in the DNA extraction step, biases in

library preparation and sequencing, and actual strain variation. Regarding the presence of polytene tissues in the DNA extraction, this could be the case of all datasets, except the ones derived from embryos. In the polytene chromosomes, most euchromatic regions fully replicate, while pericentromeric and centromeric heterochromatin does not (Gall et al. 1971). Recent results also showed that even within euchromatin, there are regions underreplicated to varying degrees, and that each polytene cell has a unique ploidy profile for these regions (Yarosh and Spradling 2014). It is also important to notice that even sequencing biases against sequences other than the 172TRs could skew their abundance estimates.

If we take the total amount of bp masked as 172TRs in the PBA (~1,8 Mbp) and the genome size estimate for *D. virilis*, it gives us a 0.48% genomic abundance. Based on the different dataset estimates, our PBA could still be missing between 14% and 73% of 172TR sequences. New 172TR abundance estimates from diploid tissues of female flies will give more precise evaluations of the PBA completeness regarding this TR.

Table 3. 172TR abundance in genomic datasets from *D. virilis*.

Dataset	Strain	DNA source	Genome proportion (%)	Monomer count [a]
Sanger raw [b]	15010-1051.87	embryos	0.55	11996
Sanger assembly	15010-1051.87	embryos	0.82	17886
PacBio raw [c]	160	adult females	0.58	12651
PacBio corrected [d]	160	adult females	0.83	18104
PacBio assembly	160	adult females	1.01	22030
Illumina	15010-1051.87	mixed adults	0.9	19630
Illumina	160	mixed adults	0.77	16795
Illumina	160	larvae	0.67	14614
Illumina	9	mixed adults	0.75	16359
Illumina	9	larvae	0.64	13959

[a] Calculated with the genome proportion and a genome size of 375.17 Mb. This is an average of genome size estimates from flow cytometry published in Bosco et al. (2007) and Gregory and Johnston (2008). [b] Random sample of 100k Sanger sequencing reads. [c] Random sample of 50k PacBio raw reads. [d] Random sample of 50k PacBio corrected reads.

172TR abundance in the *virilis* subgroup

To assess 172TR abundance in other species from the *virilis* subgroup, we used a standardized short read dataset published by Ahmed-Braimah et al. (2017). This dataset included Illumina sequencing reads from mixed males and females of *D. virilis*, *D. lummei*,

D. novamexicana and *D. americana*. Additionally, we analyzed other datasets for the strains H5 and W11 of *D. americana* (Fonseca et al. 2013), and the strains 160 and 9 of *D. virilis* (Blumenstiel 2014; Le Thomas et al. 2014). This analysis evidenced a considerable variation of 172TRs in both genome proportion and monomer counts (Table 4). The highest value of 172TR genome proportion was found in *D. americana* (2.87%), and the lowest in *D. lummei* (0.42). Because the genome proportion of a given repeat family can be skewed by shifts in genome size, we also calculated 172TR monomer counts. The highest values were still found in *D. americana* strains (36828 - 42517 copies), while the lowest values were found in *D. virilis* strains (13959 - 19630). It should be noted that, as of now, there is no published genome size estimate for *D. lummei*, precluding the calculation of 172TR monomer count in this species. In any case, even if we assume *D. lummei* to have a genome size as large as that of *D. virilis*, its monomer count would still be the lowest in the *virilis* subgroup (9161; Figure 8).

Table 4. 172TR abundance in species from the *virilis* subgroup.

Species	Strain	DNA source	Genome proportion (%)	Monomer count [a]
D. americana	H5	-	2.46	36828
D. americana	W11	-	2.38	35630
D. americana	ML97.5	mixed adults	2.84	42517
D. lummei	LM.08	mixed adults	0.42	-
D. novamexicana	15010-1031.04	mixed adults	2.31	32769
D. virilis	15010-1051.87	mixed adults	0.9	19630
D. virilis	160	mixed adults	0.77	16795
D. virilis	160	larvae	0.67	14614
D. virilis	9	mixed adults	0.75	16359
D. virilis	9	larvae	0.64	13959

[a] Monomer counts were calculated based on genome size estimates from flow cytometry published in Bosco et al. (2007) and Gregory and Johnston (2008).

The available data does not seem to point to any clear phylogenetic trend in 172TR abundance (Figure 9). Rather, the data showcase considerable variation even between recently diverged lineages (e.g. *D. americana*/*D. novamexicana* and *D. lummei*, 2.9 mya; Morales-Hojas et al. 2011). Indeed, copy number variations of TRs can be considerable even within species. For example, the alpha satellite DNA array in the human X chromosome centromere can range from 0.5 to 4.9 Mb between individuals (Miga et al. 2014). Unequal crossover is thought to be the main mechanism responsible for array size variation of complex TRs, with the highest array sizes predicted to reside in regions of lowest

recombination rates (Charlesworth et al. 1994). In *D. melanogaster*, however, recombination rate varies greatly in 100 kb windows along chromosome arms (Comeron et al. 2012). Such detailed recombination data could be useful to test the effects of recombination rate over array size in the 172TRs of *D. virilis*.

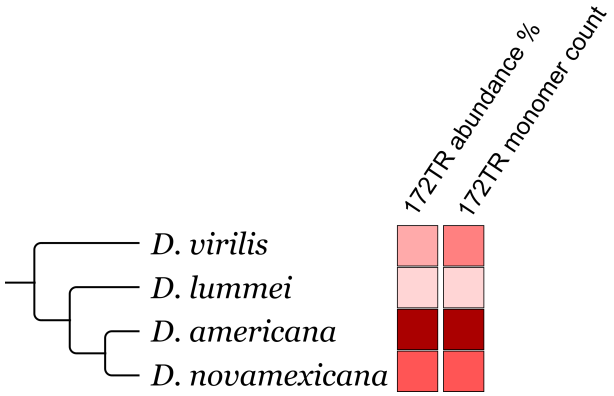


Figure 9. Phylogenetic relationships between species from the *virilis* subgroup. Genomic proportion and monomer count of the 172TRs is expressed as color intensities, i.e. lighter tones represent smaller values.

It should also be noted that the sequencing data used in this abundance comparison comes from different DNA sources, such as larvae and mixed sex adults. If we assume equal amounts of males and females in both the larvae and adult datasets, we can rely on a ¾ representation of the X chromosome. Chromosome mapping in *D. virilis* indicate the presence of 172TR arrays on the X chromosome, but not the Y (Figure 3). The smaller ratio of X chromosomes in these datasets indicates that our values are all underestimates for the 172TR abundance.

Conclusions

The last ten years were marked by a fundamental shift in the mainstream sequencing technologies. From the costly and low-throughput, but modest read sizes of Sanger capillary sequencing, to the ever cheaper and massively parallel, but very short read sizes of next generation sequencing. This shift was accompanied by a likewise decrease in assembly contiguity. It is safe to assume that 'telomere-to-telomere' genome assemblies are still far in the horizon, if they will exist at all. In any case, the development of third generation sequencing technologies brings a much welcome improvement in the quality of euchromatic genome assemblies, as exemplified by the assembly of the 172TR minisatellite. These ultra-long reads can now bridge entire TR arrays and transposable elements, allowing a better view of genome organization and enabling genomic analysis of euchromatic TRs.

In this paper, we characterized an unusually large and abundant minisatellite sequence from the *D. virilis* genome. Because micro and minisatellites are mainly used for individual identification and population genetics, they are usually only studied in the context of individual loci. Besides that, because these sequences often have low complexity, sequence similarity between different loci cannot be readily taken as indicative of homology. The 172TR minisatellite of *D. virilis* possesses a complex motif, and the hundreds of arrays spread in the genome are doubtlessly homologous. This minisatellite family is widespread in the species of the *virilis* subgroup, where it shows considerable fluctuations in copy number. This constitutes an astounding example of the dynamics of DNA repeats in eukaryotes and raises many interesting questions regarding the mobility of such repeats and their impact upon nearby genomic elements and the general chromosome stability. These and other questions are now more likely to be answered with the new PacBio assembly of *D. virilis*.

References

Abdurashitov, M. A., Gonchar, D. A., Chernukhin, V. A., Tomilov, V. N., Tomilova, J. E., Schostak, N. G., ... & Degtyarev, S. K. (2013). Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC genomics, 14(1), 771.

Ahmed-Braimah, Y. H., Unckless, R. L., & Clark, A. G. (2017). Evolutionary dynamics of male reproductive genes in the Drosophila virilis subgroup. G3: Genes, Genomes, Genetics, 7(9), 3145-3155.

Allen, S. L., Delaney, E. K., Kopp, A., & Chenoweth, S. F. (2017). Single-molecule sequencing of the *Drosophila* serrata genome. G3: Genes, Genomes, Genetics, 7(3), 781-788.

Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., ... & Tuschl, T. (2003). The small RNA profile during *Drosophila* melanogaster development. Developmental cell, 5(2), 337-350.

Armour, J. A., & Jeffreys, A. J. (1992). Biology and applications of human minisatellite loci. Current opinion in genetics & development, 2(6), 850-856.

Ashburner, M. (1989). *Drosophila*. A laboratory handbook. Cold Spring Harbor Laboratory Press.

Ashburner, M., & Bergman, C. M. (2005). *Drosophila* melanogaster: a case study of a model genomic sequence and its consequences. Genome Research, 15(12), 1661-1667.

Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., ... & Arnaiz, O. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature, 444(7116), 171.

Baimai, V. (1977). Chromosomal polymorphisms of constitutive heterochromatin and inversions in *Drosophila*. Genetics, 85(1), 85-93.

Baker M (2012). De novo genome assembly: what every biologist should know. Nature Methods, 9:333-337.

Bergman, C. M., Pfeiffer, B. D., Rincón-Limas, D. E., Hoskins, R. A., Gnirke, A., Mungall, C. J., ... & Stapleton, M. (2002). Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome biology, 3(12), research0086-1.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature biotechnology, 33(6), 623-630.

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E., & Barucca, M. (2015). Transcription of tandemly repetitive DNA: functional roles. Chromosome research, 23(3), 463-477.

Blumenstiel, J. P. (2014). Whole genome sequencing in *Drosophila virilis* identifies Polyphemus, a recently activated Tc1-like transposon with a possible role in hybrid dysgenesis. Mobile DNA, 5(1), 6.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell, 128(6), 1089-1103.

Castel, S. E., & Martienssen, R. A. (2013). RNA interference (RNAi) in the nucleus: roles for small RNA in transcription, epigenetics and beyond. Nature Reviews. Genetics, 14(2), 100.

Charlesworth B, Sniegowski P, Stephan W (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature, 371:215-220.

Cohen S, Segal D (2009). Extrachromosomal Circular DNA in Eukaryotes: Possible Involvement in the Plasticity of Tandem Repeats. Cytogenet Genome Res, 124:327–338.

Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., ... & Di Pierro, E. A. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature genetics, 49(7), 1099-1106.

Drosophila 12 Genomes Consortium. Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., ... & Pollard, D. A. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. Nature, 450(7167), 203.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., ... & Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one, 7(11), e47768.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L., & Levine, D. (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome biology and evolution, 1, 205-220.

Fonseca NA, Morales-Hojas R, Reis M, Rocha H, Vieira CP, Nolte V, Schlötterer C, Vieira J (2013). *Drosophila americana* as a model species for comparative studies on the molecular basis of phenotypic variation. Genome biology and evolution, 5(4), 661-679.

Gall JG, Cohen EH, Polan ML (1971). Repetitive DNA sequences in *Drosophila*. Chromosoma, 33:319-344.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., ... & Miller, W. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome research, 15(10), 1451-1455.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology, 11(8), R86.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological reviews, 76(1), 65-101.

Gregory TR, & Johnston JS (2008). Genome size diversity in the family Drosophilidae. Heredity, 101(3), 228-238.

Gubenko, I. S., & Evgen'ev, M. B. (1984). Cytological and linkage maps of *Drosophila virilis* chromosomes. Genetica, 65(2), 127-139.

Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Jeffreys, A. J. (1987). Highly variable minisatellites and DNA fingerprints.

Kim, V. N., Han, J., & Siomi, M. C. (2009). Biogenesis of small RNAs in animals. Nature reviews Molecular cell biology, 10(2), 126-139.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., ... & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology, 30(7), 693-700.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., ... & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome biology, 14(9), R101.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), 722-736.

Le Thomas, A., Marinov, G. K., & Aravin, A. A. (2014). A transgenerational process defines piRNA biogenesis in *Drosophila virilis*. Cell reports, 8(6), 1617-1623.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., & Karlsson, E. K. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature, 438(7069), 803.

Mahan JT, Beck ML (1986). Heterochromatin in mitotic chromosomes of the Virilis species group of *Drosophila*. Genetica, 68: 113-118.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, ... & Chan SW (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol, 14(1), R10.

Morales-Hojas R, Reis M, Vieira CP, Vieira J (2011). Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. Mol Phylogenet Evol 60(2):249–258

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., ... & Sandstrom, R. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics, 28(14), 1919-1920.

Richard GF, Kerrest A, Dujon B (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol Biol Rev. 72(4):686-727.

Rozhkov, N. V., Aravin, A. A., Zelentsova, E. S., Schostak, N. G., Sachidanandam, R., McCombie, W. R., ... & Evgen'ev, M. B. (2010). Small RNA-based silencing strategies for transposons in the process of invading *Drosophila* species. Rna, 16(8), 1634-1645.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human molecular genetics, 19(R2), R227-R240.

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O'Grady, P. M., ... & Edwards, K. (2008). Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics, 179(3), 1601-1655.

Tarailo-Graovac M, Chen N (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, 4-10.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. Genetics, 13(1), 36.

Yarosh, W., & Spradling, A. C. (2014). Incomplete replication generates somatic DNA alterations within *Drosophila* polytene salivary gland cells. Genes & development, 28(16), 1840-1855.

Walsh, J. B. (1987). Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics, 115(3), 553-567.

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marçais, G., ... & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research, 27(5), 787-792.

Discussão

Drosophila virilis é uma espécie de importância histórica no estudo de DNAs repetitivos. Por exemplo, foram trabalhos realizados em *D. virilis* que permitiram concluir que os DNA satélites não são replicados nos cromossomos politênicos das glândulas salivares de *Drosophila* (Gall et al. 1971). Foi também com a análise dos DNA satélites de *D. virilis* e de espécies próximas que se demonstrou pela primeira vez que DNA satélites podem ser compartilhados entre espécies (Gall et al. 1974). Apesar da importância deste organismo como modelo para o estudo de DNAs repetitivos, não houve muitas adições à caracterização de repetições no genoma desta espécie nos últimos trinta anos. Desta forma, os dados gerados por este projeto representam avanços significativos na caracterização de DNAs repetitivos em *D. virilis* através da integração de dados genômicos e citogenética molecular.

No capítulo 1 descrevemos o fenômeno de surgimento de DNA satélites a partir do *Helitron DINE-TR1* independentemente em duas espécies de *Drosophila*. Este é o único relato do tipo em eucariotos até o momento e o terceiro caso de relação evolutiva entre transposons e DNAs satélites em *D. virilis* (Heikkinen et al. 1995; Dias et al. 2014). Estes três casos distintos descritos apenas em *D. virilis* e o número crescente de outros exemplos na literatura sugerem que o surgimento de DNA satélites a partir de TEs possa ser uma via recorrente na evolução de DNAs repetitivos (Meštrović et al. 2015).

A caracterização deste DNA satélite derivado do *Helitron DINE-TR1* em *D. virilis* também demonstrou que, ao contrário do suposto por Melters et al. (2013), esta sequência não possui localização centromérica, com exceção de um par cromossômico (fig. 5 do Cap. 1). Em seu trabalho, Melters et al. (2013) assumiu que a repetição em tandem mais abundante em um genoma corresponderia ao DNA satélite centromérico. Em um trabalho recente, Bilinski et al. (2017) testaram esta hipótese em 16 espécies de plantas e na maioria delas a hipótese também se mostrou falsa. Estes dados ilustram como a citogenética é essencial na análise de genomas, reforçando que análises bioinformáticas por si só não são suficientes para determinar a distribuição genômica de DNAs repetitivos.

Na análise do *DINE-TR1* no genoma de *D. virilis*, notamos vários casos em que cópias do transposon inteiro se encontravam organizadas em tandem (fig. 1 do Cap. 1). Este tipo de fenômeno já havia sido observado em outros organismos, porém em menor escala. No segundo capítulo, invocamos um mecanismo alternativo de transposição de *Helitrons* para explicar esse fenômeno de inserções em tandem (fig. 3B do Cap. 2). O chamado "modelo sequencial" de transposição discutido neste capítulo postula a existência de intermediários circulares de *Helitrons*, algo não demonstrado até então. Quase simultaneamente à publicação do capítulo 2, foi publicado um trabalho que demonstrava experimentalmente a formação de círculos extra-cromossômicos durante a transposição de um *Helitron* do

morcego *Myotis lucifugus* em cultura de células (Grabundzija et al. 2016). Esta nova evidência reforça o modelo sequencial de transposição de *Helitrons* e indica que este possa ser de fato um mecanismo para geração de cópias em tandem destes elementos. Por outro lado, dados recentes demonstram a existência de inserções em tandem de TEs de todas as classes em *D. melanogaster* e sugerem que a preferência por sítios de inserção específicos possa explicar este fenômeno (McGurk & Barbash 2017). As duas hipóteses não são excludentes e mais ensaios de transposição de *Helitrons*, além de análises genômicas similares a esta em outras espécies, podem ajudar a determinar a importância relativa de cada mecanismo na geração de cópias em tandem.

No último capítulo, caracterizamos um minissatélite compartilhado entre as espécies do subgrupo *virilis*. Esta repetição é extraordinária no sentido de possuir monômeros grandes, normalmente só presentes em DNAs satélites, e apresentar alta abundância e dispersão no genoma. Como os minissatélites foram historicamente estudados como marcadores para identificação de indivíduos e populações, eles são quase sempre analisados no contexto de loci específicos, e não em um contexto genômico e/ou evolutivo (Armour & Jeffreys 1992; Tautz 1993). Pelo mesmo motivo, toda a literatura de minissatélites apresenta um viés de caracterização favorecendo os loci com tamanho mais variável entre indivíduos (maior número de alelos) e tamanho total não muito grande, de modo que sejam manejáveis com as técnicas comumente utilizadas para *screening* (geralmente PCR). Por estes motivos, as características atribuídas a minissatélites podem representar apenas uma fração da diversidade real destas repetições, e casos como o minissatélite 172TR de *D. virilis* podem ser mais comuns do que se imagina. De fato, existe pelo menos mais um caso em *Drosophila* de repetições em tandem eucromáticas que poderiam ser classificadas como um minissatélite atípico (Nozawa et al. 2006).

Um outro questionamento fundamental surge quando se analisa este tipo de repetição eucromática: como estas cadeias se movem? Ao contrário dos transposons, as repetições em tandem não codificam a maquinaria necessária para mediar sua mobilização. Entretanto, ao visualizar a amplitude da dispersão genômica da família 172TR (fig. 3 do Cap. 3), fica muito claro que estas sequências são, ou já foram, móveis. O modelo mais completo para explicar a mobilidade de cadeias de repetições em tandem postula a formação de círculos extra-cromossômicos circulares através da recombinação homóloga entre monômeros na mesma cadeia. Estes círculos poderiam então se reintegrar no genoma através de recombinação ilegítima (Cohen & Segal 2009). A primeira parte do modelo já foi demonstrada experimentalmente (Cohen et al. 2003). Entretanto, a integração destas repetições de volta no genoma ainda não foi demonstrada. Usando a nova montagem PacBio do genoma de *D. virilis* como referência, talvez seja possível identificar inserções polimórficas das cadeias 172TR utilizando métodos de detecção de inserções de transposons.

Estes métodos utilizam *reads* curtas de sequenciamento e um genoma de referência com TEs anotados para identificar polimorfismos de inserção de TEs em relação à referência (Nelson et al. 2017). Já existem alguns conjuntos de dados de *reads* curtas para linhagens distintas de *D. virilis* (Tabela 3, Cap. 3). Esta análise pode ser promissora para elucidar os passos da integração de repetições em tandem e até mesmo para estimar a frequência com que tais eventos ocorrem.

Conclusões

- Eventos de surgimento de DNAs satélites a partir de TEs parecem ser frequentes no genoma de *D. virilis*, com três eventos distintos já relatados envolvendo TEs de diferentes classes;
- O modelo sequencial de transposição de *Helitrons* pode ajudar a explicar as frequentes inserções em tandem destes elementos observadas no genoma de *D. virilis* e outros eucariotos;
- Minissatélites podem possuir monômeros consideravelmente complexos, com alta abundância e dispersão genômica;
- A utilização de dados de sequenciamento de terceira geração pode contribuir com uma melhoria significativa na contiguidade do genoma, principalmente nas regiões repetitivas eucromáticas;
- Apesar de *D. virilis* ter sido objeto de pesquisas por quase um século e ter seu genoma publicamente disponível há dez anos, a caracterização dos DNAs repetitivos deste genoma ainda se encontra em sua fase inicial.

Referências

Abdurashitov, M. A., Gonchar, D. A., Chernukhin, V. A., Tomilov, V. N., Tomilova, J. E., Schostak, N. G., ... & Degtyarev, S. K. (2013). Medium-sized tandem repeats represent an abundant component of the *Drosophila virilis* genome. BMC genomics, 14(1), 771.

Armour, J. A., & Jeffreys, A. J. (1992). Biology and applications of human minisatellite loci. Current opinion in genetics & development, 2(6), 850-856.

Bilinski, P., Han, Y., Hufford, M. B., Lorant, A., Zhang, P., Estep, M. C., ... & Ross-Ibarra, J. (2017). Genomic abundance is not predictive of tandem repeat localization in grass genomes. *PloS one*, *12*(6), e0177896.

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E., & Barucca, M. (2015). Transcription of tandemly repetitive DNA: functional roles. Chromosome research, 23(3), 463-477.

Bosco, G., Campbell, P., Leiva-Neto, J. T., & Markow, T. A. (2007). Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics, 177(3), 1277-1290.

Brajkovic J, Feliciello I, Bruvo-MadWaric B, Ugarkovic D (2012). Satellite DNA-Like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. G3 (Bethesda) 2:931–941.

Charlesworth B, Sniegowski P, Stephan W (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature, 371:215-220.

Cohen, S., Yacobi, K., & Segal, D. (2003). Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. *Genome research*, *13*(6a), 1133-1145.

Cohen S, Segal D (2009). Extrachromosomal Circular DNA in Eukaryotes: Possible Involvement in the Plasticity of Tandem Repeats. Cytogenet Genome Res, 124:327–338.

Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM, & FlyBase Consortium (2007). FlyBase: genomes by the dozen. Nucleic acids research, 35(suppl 1):D486-D491.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genetics 7(12): e1002384.

de Souza FS, Franchini LF, Rubinstein M (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol, 30(6):1239-1251.

Dias, G. B., Svartman, M., Delprat, A., Ruiz, A., & Kuhn, G. C. (2014). Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of *Drosophila virilis*. Genome biology and evolution, 6(6), 1302-1313.

Dotto BR, Carvalho EL, Freitas A, Duarte LF, Pinto PM, Ortiz MF, Wallau GL (2015). HTT-DB-Horizontally transferred transposable elements database. Bioinformatics, btv281.

Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry. Part A: the journal of the International Society for Analytical Cytology, 51(2), 127-8.

Dover GA (1982). Molecular drive: a cohesive mode of species evolution. Nature 299:111-117.

Dover, G. (2002). Molecular drive. Trends in Genetics, 18(11), 587-589.

Drosophila 12 Genomes Consortium. Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., ... & Pollard, D. A. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. Nature, 450(7167), 203.

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome biology and evolution, 1:205.

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

Gall JG, Cohen EH, Polan ML (1971). Repetitive DNA sequences in *Drosophila*. Chromosoma, 33:319-344.

Gall, J. G. (1973). Repetitive DNA in *Drosophila*. In Molecular Cytogenetics (pp. 59-74). Springer New York.

Gall, J. G., Cohen, E. H., & Atherton, D. D. (1974). The satellite DNAs of *Drosophila virilis*. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 38, pp. 417-421). Cold Spring Harbor Laboratory Press.

Gatti M, & Pimpinelli S (1992). Functional elements in *Drosophila melanogaster* heterochromatin. Annual review of genetics, 26(1):239-276.

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual review of genetics, 44, 445-477.

Gong, Z., Wu, Y., Koblížková, A., Torres, G. A., Wang, K., Iovene, M., ... & Macas, J. (2012). Repeatless and repeat-based centromeres in potato: implications for centromere evolution. The Plant Cell, 24(9), 3559-3574.

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

Gregory, T. R. (2005). Genome size evolution in animals. The evolution of the genome, 1, 4-87.

Gregory TR, & Johnston JS (2008). Genome size diversity in the family Drosophilidae. Heredity, 101(3), 228-238.

Heikkinen E, Launonen V, Müller E, Bachmann L. The pvB370 BamHI Satellite DNA Family of the *Drosophila virilis* Group and Its Evolutionary Relation to Mobile Dispersed Genetic pDv Elements. J Mol Evol, 41:604-614, 1995.

Hourcade, D., Dressler, D., & Wolfson, J. (1973). The amplification of ribosomal RNA genes involves a rolling circle intermediate. Proceedings of the National Academy of Sciences, 70(10), 2926-2930.

Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P (2011). The struggle for life of the genome's selfish architects. Biol Direct, 6: 19.

Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends in Genetics.

Kapitonov VV, Holmquist GP, Jurka J (1998). L1 repeat is a basic unit of heterochromatin satellites in cetaceans. Molecular biology and evolution, 15(5):611–612.

Kapitonov VV, Jurka J (2001). Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98:8714–8719.

Kapitonov VV, Jurka J (2007). *Helitrons* on a roll: eukaryotic rolling-circle transposons. Trends in Genetics; 23(10):521-529.

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012). The 1.688 Repetitive DNA of *Drosophila*: Concerted Evolution at Different Genomic Scales and Association with Genes. Mol Biol Evol, 29(1):7-11.

Kuhn, G. C. S. (2015). 'Satellite DNA transcripts have diverse biological roles in *Drosophila*'. Heredity, 115(1), 1.

Lee YCG (2015). The Role of piRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in *Drosophila melanogaster*. PLOS Genet 11(6), e1005269.

Liao D (1999). Concerted evolution: molecular mechanism and biological implications. The American Journal of Human Genetics, 64(1):24-30.

Macas J, Koblížková A, Navrátilová A, Neumann P (2009). Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448: 198-206.

Mahan JT, Beck ML (1986). Heterochromatin in mitotic chromosomes of the *Virilis* species group of *Drosophila*. Genetica, 68: 113-118.

Markow T (2015). The secret lives of *Drosophila* flies. eLife 4:e06793.

Martiessen RA (2003). Maintenance of heterochromatin by RNA interference of tandem repeats. Nature Genetics, 35:213-214.

Masumoto H, Nakano M, Ohzeki J (2006). The role of CENP-B and α-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. Chromosome Res., 12(6): 543-556.

Maumus F, Fiston-Lavier AS, Quesneville H (2015). Impact of transposable elements on insect genomes and biology. Current Opinion in Insect Science, 7:30-36.

McGurk, M. P., & Barbash, D. A. (2017). Continuous generation of tandem transposable elements in *Drosophila* populations provides a substrate for the evolution of satellite DNA. *bioRxiv*, 158386.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, ... & Chan SW (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol, 14(1), R10.

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., & Plohl, M. (2015). Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome research*, *23*(3), 583-596.

Nelson, M. G., Linheiro, R. S., Bergman, C. M. (2017). McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3: Genes, Genomes, Genetics, 7(8), 2763-2778.

Nozawa, M., Kumagai, M., Aotsuka, T., & Tamura, K. (2006). Unusual evolution of interspersed repeat sequences in the *Drosophila* ananassae subgroup. *Molecular biology and evolution*, *23*(5), 981-987.

Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, Raimondi E, Giulotto E (2010). Uncoupling of satellite DNA and centromeric function in the genus Equus. PLoS Genet, 6(2), e1000.

Plohl M, Luchetti A, Meštrović N, Mantovani B (2008). Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. Gene, 409(1):72-82.

Plohl, M., Meštrović, N., & Mravinac, B. (2014). Centromere identity from the DNA point of view. Chromosoma, 123(4), 313-325.

Richard GF, Kerrest A, Dujon B (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol Biol Rev. 72(4):686-727.

Satovic E, Plohl M (2013) Tandem repeat-containing MITEs in the clam *Donax trunculus*. Genome Biol Evol, 5(12):2549–2559.

Sentmanat MF, Elgin SC (2012). Ectopic assembly of heterochromatin in *Drosophila* melanogaster triggered by transposable elements. Proc Natl Acad Sci USA 109(35):14104–14109.

Sharma A, Wolfgruber TK, Presting GG (2013). Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14:142.

Smith JDL, Bickham JW, Gregory TR (2013). Patterns of genome size diversity in bats (order Chiroptera). Genome 56:457-472.

Strachan T, Read A. Organization of the Human Genome. In: Human Molecular Genetics, 4th edition. Garland Science, 2010.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nature Genetics, 43:1160-1163.

Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In DNA fingerprinting: State of the science (pp. 21-28). Birkhäuser Basel.

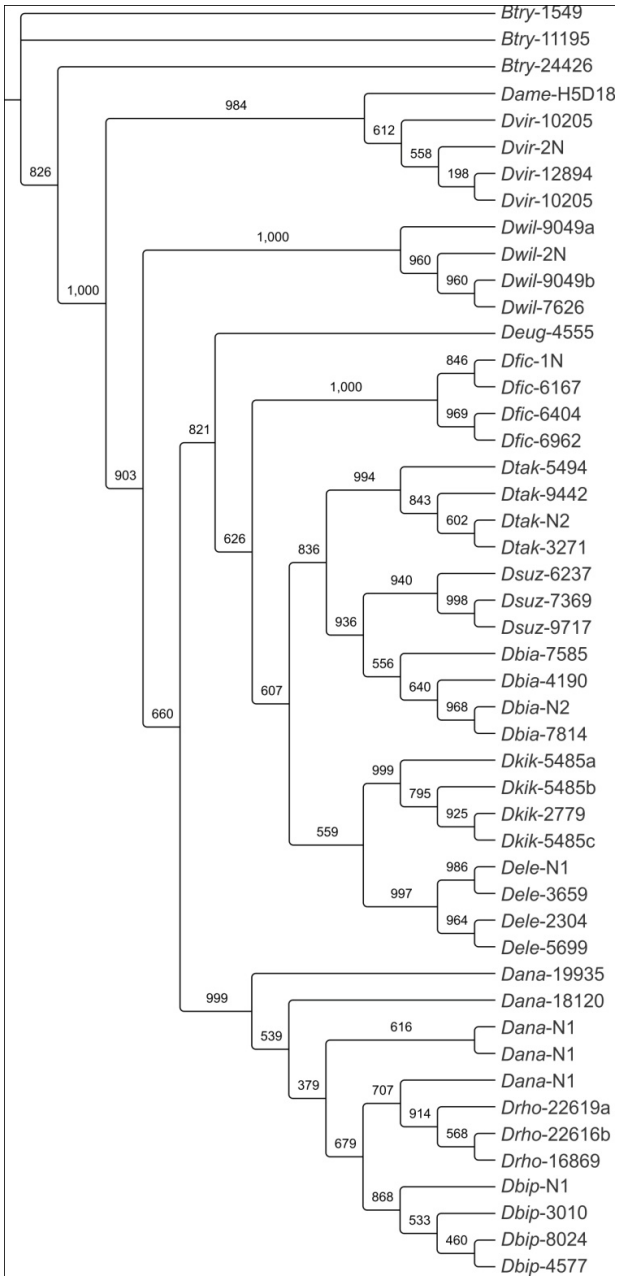
Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. Genetics, 13(1), 36.

Ugarkovic D (2005). Functional elements residing within satellite DNAs. EMBO J. 6(11): 1035-1039.

Westerman M, Barton NH, Hewitt GM (1987). Differences in DNA content between two chromosomal races of the grasshopper *Podisma pedestris*. Heredity 58:221-228.

Anexos

Anexo 1. Material suplementar do capítulo 1



Supplementary Figure 1. Maximum Likelihood phylogeny of *DINE-TR1* sequences. The tree was constructed with PhyML (Guindon and Gascuel 2003) based on an nucleotide alignment of a sample of *DINE-TR1* sequences that include the entire block A (5' end) plus the first 150bp tandem repeat. The substitution model used was TPM1uf+G (alpha fixed in 1.24). Values above the nodes represent the bootstrap support after 1000 replicates. The consensus *DINE-TR1* from Repbase are included for some species. Species names are indicated by the three initials followed by the contig number they were extracted from. Full names are on Figure 2 from the main text.

Supplementary Table 1. Significant BLAST hits using *DINE-TR1* as query and excluding *Drosophila* from the nr/nt database in GenBank. (Last accessed in January 2015)

Accession nº	Description	Query cover (%)	Identity (%)	E value
AJ535756.1	*Bactrocera oleae* ovoA gene for zinc finger transcription factor	74%	69%	1e-24
AJ535757.1	*Bactrocera oleae* ovoB gene for zinc finger transcription factor	74%	69%	1e-24
FJ710563.1	*Bactrocera dorsalis* clone Bdor_pnrfos3.scaffold_0 pnr gene, partial sequence	80%	67%	2e-16
D89934.1	*Musca domestica* DNA, mariner-like element	8%	85%	6e-15
XM_011210380.1	PREDICTED: *Bactrocera dorsalis* band 4.1-like protein 5 (LOC105229888), mRNA	77%	65%	3e-13
JX315619.1	*Cochliomyia macellaria* transformer (tra) gene, complete cds	12%	78%	3e-13
HQ609500.1	*Lucilia cuprina* HSP24 gene, complete cds	8%	83%	3e-13
AJ297850.1	*Musca domestica* partial bcd gene for bicoid protein, promoter region, exon 1 and joined CDS	8%	86%	3e-13
JX996042.1	*Stomoxys calcitrans* Orco (Orco) gene, partial cds	13%	74%	4e-11
AF182164.1	*Stomoxys calcitrans* defensin 2a (smd2a) gene, complete cds	8%	82%	4e-11
AF283258.1	*Musca domestica* LPR cytochrome P450 (CYP6D3) gene, CYP6D3v3 allele, 5' flanking region and partial cds	11%	76%	1e-10
AF200191.1	*Musca domestica* cytochrome P450 6D3 (CYP6D3) gene, CYP6D3-v1 allele, complete cds; and cytochrome P450 6D1 (CYP6D1) gene, CYP6D1-v6 allele, partial cds	11%	75%	2e-08
NM_001287094.1	*Musca domestica* phormicin-like (defensin-1), mRNA	9%	77%	3e-07
FJ710559.1	*Bactrocera dorsalis* clone Bdor_pnrfos2.scaffold_0 pnr (pnr) gene, partial cds	42%	79%	3e-07
FJ710555.1	*Bactrocera dorsalis* clone Bdor_pnrfos1.scaffold_0 pnr (pnr) gene, partial cds	42%	79%	3e-07
EF629377.1	*Haematobia irritans* clone HI03 microsatellite sequence	14%	71%	3e-07
AC182713.2	*Populus trichocarpa* clone Pop1-86K12, complete sequence	7%	81%	3e-06

AY256681.1	*Musca domestica* hexamerin F1 (HexF1) gene, complete cds	14%	70%	3e-06
XM_011207367.1	PREDICTED: *Bactrocera dorsalis* cationic amino acid transporter 4 (LOC105227831), mRNA	20%	67%	1e-05
AM940018.1	*Glossina morsitans* morsitans attA11 gene, attB gene, attA12 gene, attA21 gene and attD gene, clone 39G22	10%	75%	4e-05
XM_011212522.1	PREDICTED: *Bactrocera dorsalis* uncharacterized LOC105231306 (LOC105231306), transcript variant X2, mRNA	12%	71%	1e-04
XM_011212521.1	PREDICTED: *Bactrocera dorsalis* uncharacterized LOC105231306 (LOC105231306), transcript variant X1, mRNA	12%	71%	1e-04
XM_011210542.1	PREDICTED: *Bactrocera dorsalis* peptidylprolyl isomerase domain and WD repeat-containing protein 1 (LOC105229994), transcript variant X3, mRNA	42%	67%	1e-04
XM_011210541.1	PREDICTED: *Bactrocera dorsalis* peptidylprolyl isomerase domain and WD repeat-containing protein 1 (LOC105229994), transcript variant X1, mRNA	42%	67%	1e-04
EU189083.1	*Chymomyza costata* timeless gene, complete cds	54%	65%	1e-04

Supplementary Table 2. Significant BLAST hits of *DINE* insertions near or at genes from *D. virilis*. (nr/nc database for *D. virilis* last accessed in January 2015)

Accession nº	Description	Query Cover (%)	Identity (%)	E value
M34544.1	*D.virilis* sevenless gene, exon 3,4,5,6 and 7	100	94	0.0
AF098329.1	*Drosophila virilis* eyeless protein gene; exons 2 and 3, partial sequence; intron 2, complete sequence; and partial cds	92	86	0.0
AY665299.1	*Drosophila virilis* strain 160 hsp70d-hsp70e intergenic region, complete sequence	100	84	0.0
AY333070.1	*Drosophila virilis* antennapedia (Antp), CG10013 (CG10013), CG31217 (CG31217), and ultrabithorax (Ubx) genes, complete cds	100	78	3e-148
AB271538.1 AB986232.1 AB986231.1 AB986229.1 AB986228.1 AB080189.1	*Drosophila virilis* Acph gene for acid phosphatase, complete cds (accession numbers from different populations)	81*	73*	1e-46*
AY186999.1	*Drosophila virilis* unc-5 (unc-5) gene, partial cds; and ap (ap), vlc (vlc), CG17337 (CG17337), and CG13533 (CG13533) genes, complete cds	97	73	1e-46
AB080189.1	*Drosophila virilis* Acph gene for acid phosphatase, complete cds	69	73	1e-46
AF045585.1	*Drosophila virilis* SISA (sisA) gene, complete cds	66	76	1e-44
AB986230.1	*Drosophila virilis* gene for acid phosphatase, complete cds, strain: Acph-2, country: Japan:Horioka	72	84	6e-31
M35372.1	*Drosophila virilis* rough gene	64	67	7e-18
L37035.1	*Drosophila virilis* brown protein (bw) gene, complete cds	40	71	4e-15
AY128944.1	*Drosophila virilis* yellow (y) gene, complete cds	72	77	4e-14
AF190404.1	*Drosophila virilis* Prospero (pros) gene, upstream genomic sequence	50	86	1e-07

*Values given are from the best hit to the Acph gene.

Supplementary Table 3. A sample of contigs entirely covered by the CTRs of *DINE-TR1* in *D. virilis* and *D. biarmipes*.

	Accession number	Contig number	Contig size (bp)
D. virilis	AANI01000001.1	0	4.376
	AANI01000004.1	3	4.066
	AANI01000022.1	21	1.021
	AANI01000048.1	47	1.606
	AANI01000084.1	83	3.549
	AANI01000093.1	92	1.083
	AANI01006142.1	6159	2.237
	AANI01006670.1	6687	1.504
	AANI01006678.1	6695	5.614
	AANI01007101.1	7119	2.032
D. biarmipes	AFFD02000580.1	581	1.331
	AFFD02000704.1	705	1.481
	AFFD02000901.1	902	2.486
	AFFD02001852.1	1853	1.075
	AFFD02004922.1	4928	1.347
	AFFD02005492.1	5499	3.015
	AFFD02005514.1	5521	2.890
	AFFD02005548.1	5555	3.961
	AFFD02005597.1	5604	2.076
	AFFD02005625.1	5632	7.582

Supplementary Table 4. *DINE-TR1* and total *Helitron* proportions in the piRNA clusters defined in *D. virilis* by Rozhkov et al. (2010).

Cluster ID	Chromosome /region[a]	Coordinates	Cluster size (Kb)	*DINE-TR1* proportion (%)[b]	Helitron proportion (%)[b]
1	2/T	>scaffold_12822: 2351-38251	36	3,2	4,6
2	6/T	>scaffold_13052: 1691054-1826988	136	1,1	8,2
3	4/N	>scaffold_12723: 984749-1045698	61	1,8	2,4
4	3/T	>scaffold_10322: 421703-452327	31	6,3	10,8
5	3/N	>scaffold_13049: 24525546-24741236	216	2,7	4,6
6	-	>scaffold_12100: 3319-269238	266	7,6	11,8
7	5/N	>scaffold_13324: 1490374-1800556	311	4,1	8,9
8	-	>scaffold_12799: 2528-283391	281	2,0	8,9
9	-	>scaffold_12937: 108029-258729	151	2,8	11,3
10	5/T	>scaffold_12823: 174998-195392	21	0	2,8
11	X	>scaffold_12932: 676403-707267	31	0	0,5
12	2/N	>scaffold_12954: 311677-581831	271	0,9	6,8
13	3/N	>scaffold_13049: 24340946-24411205	71	1,4	3,1
14	5/N	>scaffold_13324: 55438-430851	376	1,4	6,7
15	-	>scaffold_12958: 586817-757593	171	1,0	9,4
16	2/N	>scaffold_12954: 605843-776563	171	0	10,6
17	-	>scaffold_12734: 387869-503389	116	1,4	7,5
18	2	>scaffold_12855: 9711211-9806955	96	0,4	4,5
19	3/N	>scaffold_13049: 25145270-25268649	123	9,2	12,1
20	5/N	>scaffold_13324: 1265331-1441221	176	2,1	6,0

[a]Chromosome regions identified as N for nearcentromere or T for telomere (as in Rozhkov et al. 2010).
[b]Proportions were calculated after repeat masking the clusters with the CENSOR tool (Kohany et al. 2006).

Anexo 2. Lista de publicações

Publicações como primeiro autor

Dias, G. B., Heringer, P., Svartman, M., & Kuhn, G. C. (2015). *Helitrons* shaping the genomic architecture of *Drosophila*: enrichment of *DINE-TR1* in α-and β-heterochromatin, satellite DNA emergence, and piRNA expression. Chromosome research, 23(3), 597-613.

Dias, G. B., Heringer, P., & Kuhn, G. C. (2016). *Helitrons* in *Drosophila*: Chromatin modulation and tandem insertions. Mobile genetic elements, 6(2), e1154638.

Araújo, N. P., **Dias, G. B.***, Amaro, B. D., Kuhn, G. C. S., & Svartman, M. (2016). The complete mitochondrial genomes of two Atlantic spiny rats, genus Trinomys (Rodentia: Echimyidae), from low-pass shotgun sequencing. Gene Reports, 5, 18-22.

* Autoria compartilhada

Publicações como coautor

Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas, F., Puig, M., Casillas, S., Ràmia, M., Egea, R., Negre, B., Mir, G., Camps, J., Moncunill, V., Ruiz-Ruano, J., Cabrero, J., Lima, L. G., **Dias, G. B.**, Ruiz, J. C., Kapusta, A., Garcia-Mas, J., Gut, M., Gut, I. G., Torrents, D., Camacho, J. P., Kuhn, G. C. S., Feschotte, C., Clark, A. G., Betrán, E., Barbadilla, A., Ruiz, A. (2014). Genomics of ecological adaptation in cactophilic *Drosophila*. Genome biology and evolution, 7(1), 349-366.

Araújo, N. P., de Lima, L. G., **Dias, G. B.**, Kuhn, G. C. S., de Melo, A. L., Yonenaga-Yassuda, Y., Stanyon, R., Svartman, M. (2017). Identification and characterization of a subtelomeric satellite DNA in Callitrichini monkeys. DNA Research, 24(4), 377-385.

Palacios-Gimenez, O. M., **Dias, G. B.**, Lima, L. G., Kuhn, G. C. S., Ramos, E., Martins, C., Cabral-de-Mello, D. C. (2017). High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*. Scientific Reports, 7.