UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

DEPARTAMENTO DE BIOLOGIA GERAL

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Leonardo Gomes de Lima

DISSECANDO O SATELITOMA DE ESPÉCIES DE *DROSOPHILA* COM GENOMAS SEQUENCIADOS

Belo Horizonte

2017

Leonardo Gomes de Lima

DISSECANDO O SATELITOMA DE ESPÉCIES DE *DROSOPHILA* COM GENOMAS SEQUENCIADOS

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

Orientador: Prof. Dr.Gustavo Campos e Silva Kuhn
Coorientadora: Profa. Dra. Marta Svartman

Belo Horizonte, 2017

FICHA CATALOGRÁFICA

043 Lima, Leonardo Gomes de.
Dissecando o satelitoma de espécies de *Drosophila* com genomas sequenciados [manuscrito] / Leonardo Gomes de Lima. – 2017.
201 f. : il. ; 29,5 cm.

Orientador: Prof. Dr.Gustavo Campos e Silva Kuhn. Coorientadora: Profa. Dra. Marta Svartman.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. DNA Satélite. 3. Evolução Molecular. 4. *Drosophila*. I. Kuhn, Gustavo Campos e Silva. II. Svartman, Marta. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575

Ficha catalográfica elaborada pela Bibliotecária Fabiane C M Reis. CRB/6 - 2680

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

ATA DA DEFESA DE TESE

96/2017
entrada
2º/2013
CPF:
033.371.015-00

Leonardo Gomes de Lima

Às treze horas do dia **29 de setembro de 2017**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Dissecando o satelitoma de espécies de Drosophila com genomas seqüenciados"**, requisito para obtenção do grau de Doutor em **Genética.** Abrindo a sessão, o Presidente da Comissão, **Gustavo Campos e Silva Kuhn**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Gustavo Campos e Silva Kuhn	UFMG	26013664862	APROVADO
Maura Helena Manfrin	USP	071.746.198-00	APROVADO
Jerônimo Conceição Ruiz	Fiocruz	33110923870	APROVADO
Renan Pedra de Souza	UFMG	06449106601	APROVADO
Carlos Renato Machado	UFMG	71055886772	APROVADO

Pelas indicações, o candidato foi considerado: APROVADO
O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 29 de setembro de 2017.

Gustavo Campos e Silva Kuhn _____

Maura Helena Manfrin _____

Jerônimo Conceição Ruiz _____

Renan Pedra de Souza _____

Carlos Renato Machado _____

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

"Dissecando o satelitoma de espécies de Drosophila com genomas seqüenciados."

Leonardo Gomes de Lima

Tese aprovada pela banca examinadora constituída pelos Professores:

Gustavo Campos e Silva Kuhn
UFMG

Maura Helena Manfrin
USP

Jerônimo Conceição Ruiz
Fiocruz

Renan Pedra de Souza
UFMG

Carlos Renato Machado
UFMG

Belo Horizonte, 29 de setembro de 2017.

Dedico esta tese à minha família.

Agradecimentos

Agradecimentos são, de forma geral, desnecessários. Aqueles que merecem poucas vezes sentem necessidade de recebê-los, enquanto os que esperam por estas manifestações poucas vezes as merecem.

Agradeço às pessoas que constituem o corpo pulsante da Universidade Federal de Minas Gerais. Pessoas de todas as áreas que contribuem para a manutenção desta instituição que carinhosamente me acolheu por um longo período.

Agradeço também aos professores e alunos do Programa de Pós-Graduação em Genética. Os professores por me municiarem de conhecimento e modelos a serem seguidos. Os alunos por me deixarem representá-los por mais de dois anos e meio, além de todas as amizades conquistadas neste período (em especial o Pablo, José Eustáquio, Jean, Renata e Luciana).

Aos meus orientadores, Professor Dr. Gustavo Kuhn e Professora Dra. Marta Svartman por toda a ajuda na realização deste projeto e todo apoio nesse início de carreira científica. Também devo agradecimentos aos colegas de Laboratório que me aguentaram por todos estes anos: Guilherme, Pedro, Luísa, Mathias, Radarane, Alice e Diego.

Às amizades que de diversas formas foram essenciais para que conseguisse terminar este projeto. A essas pessoas que me mostraram que prefiro aqueles que sabem rir de seus tropeços, não se encantam com triunfos, que não se consideram eleitos antes da hora e não fogem de sua mortalidade. Dentro de todos os processos desta vida há conexões que são construídas sob os alicerces do respeito mútuo e irrestrito, embora muitas vezes não se perceba isso. Dito isto, não há amigos melhores que nossos velhos amigos, por isso, agradeço a Marquinhos, Rodrigo, Marcelinho, Steven, Maria, Goiaba e, em especial, a Thalles (papai)

que me ajudaram a sobreviver durante estes quatro anos mais conturbados que o cenário político brasileiro.

Agradeço também à minha grande amiga e chefa Naiara Araújo, pela parceria, amizade e conversas que me fizeram virar seu fã.

E ao meu grande amigo José Ricardo, pessoa de capacidade, inteligência e companheirismo singulares. Pelas incontáveis discussões científicas, filosóficas, de cunho pessoal ou banal regadas a muito café, cigarros e algumas cervejas. Aprendi muito com você e para meus parceiros “tenho a oferecer minha presença, talvez até confusa, mas leal e intensa/ No melhor Marvin Gaye, sabadão na marginal/ O que será, será, somos nós, vamos até o final”.

Aos meus companheiros de dia a dia, Jaspion e Graveto, pela companhia, ajuda em momentos difíceis, além de carinho e amor.

Em especial, agradeço à minha família, Prof. Lima, Dª Maria José, Manu, Camila e Denise. Aos meus pais pelo incentivo hercúleo de me ajudar durante os mais de dez anos longe de casa. Sem o seu apoio e carinho nada do que consegui teria sido possível. Às minhas irmãs, Manuela e Camila, por um amor incondicional e exemplos de seres humanos que devem ser exaltados sempre. Os seus cuidados, preocupações e ações (mesmo que às vezes eu não concordasse) me mostraram que gosto de caminhar perto de pessoas de verdade. O essencial faz a vida valer a pena, e para mim basta o essencial.

“A maior riqueza do homem

é sua incompletude.

Nesse ponto sou abastado.

Palavras que me aceitam como sou — eu não aceito.

Não aguento ser apenas um sujeito que abre portas, que puxa

válvulas, que olha o relógio, que compra pão

às 6 da tarde, que vai lá fora, que aponta lápis,

que vê a uva etc. etc.

Perdoai.

Mas eupreciso ser Outros.

Eu penso renovar o homem usando borboletas.”

Manoel de Barros

RESUMO

As sequências de DNA satélite estão presentes em praticamente todos os genomas eucarióticos estudados, e geralmente estão associados à heterocromatina. Além disso, análises recentes têm associado sequências de DNAs satélites a vários papéis biológicos. No entanto, estes importantes elementos genéticos foram historicamente negligenciados e atualmente ainda há uma baixa representatividade destes elementos nos bancos de dados. Ademais, poucos trabalhos têm focado uma análise de caracterização de DNAs satélites em um contexto filogenético no gênero *Drosophila*. Neste trabalho abordamos esta questão realizando uma análise do conjunto total de sequências de DNAs satélites (satelitoma) em 36 espécies do gênero *Drosophila*. Para a caracterização *de novo* das sequências de DNA satélites foi utilizado o pipeline *RepeatExplore*r, além de análises de alinhamento, transcrição e hibridização *in situ*. Aqui, descrevemos 172 famílias de DNAs satélites presente nas 36 espécies de *Drosophila*. Neste trabalho 133 sequências de DNA satélite foram caracterizadas pela primeira vez. A análise do conteúdo de repetitivo em uma aborgadem filogeneticamente ampla revelou a natureza divergente das seqüências de DNAs satélites nos genomas de *Drosophila*. De forma geral, observamos que o conteúdo de DNAs satélites variou entre 0,54% em *D. arizonae* até 27,4% do genoma de *D. montana*. Além disso, confirmamos a hipótese da biblioteca de DNAs satélites ao descrevermos a manutenção de diversas famílias DNAs satélites compartilhadas entre espécies estreitamente relacionadas. Descobrimos também que a família de DNA satélite *1.688* é compartilhada por 13 espécies divergiram há 27 milhões de anos, indicando a maior conservação de uma família de DNA satélite caracterizado em *Drosophila*. Por fim, a análise de correlação entre DNAs satélites e o tamnho do genoma de casa espécie mostrou que as alterações no tamanho do genoma da *Sophophora* estão positivamente correlacionadas com a abundância de elementos transponíveis, enquanto a variação no tamanho do genoma de espécies do subgênero de *Drosophila* está fortemente correlacionada à variação do percentual de DNAs satélites.

Palavras-Chave: DNA satélite; *Drosophila*; Evolução do Genoma

ABSTRACT

Satellite DNAs (satDNA) are ubiquitously present in eukaryotic genomes and are usually associated with heterochromatic regions of the genome. Moreover, this genetic element has been recently associated with several biological roles. However, these important genetic elements have been historically neglected and currently there is still a low representativeness of these elements in the databases. In addition, only few studies have focused on a characterization of satDNAs in a phylogenetic context in *Drosophila*. In this work, we approach this question by performing a de novo characterization of the total set of satDNA sequences (satellitome) present in the genome of 36 *Drosophila* species. We used the RepeatExplorer pipeline for the de novo identification, as well as alignment, transcription and in situ hybridization analyzes. Herein we described 172 satDNA families, being 133 of them newly described satDNA sequences. Repeat analysis within a phylogenetic framework has revealed the profound divergent nature of satDNA sequences in *Drosophila* genomes. We observed that the satDNA content varied from 0.54% of *D. arizonae* genome to 27.4% of *D. montana*. We confirmed the satDNA library hypothesis evidencing the maintenance of satDNA families shared among closely related species. We also described that the *1.688* satDNA family is present in 13 species of group melanogaster which diverged ~27 Mya, indicating the most striking conservation of a satDNA family in *Drosophila*. Finally, we found that changes in genome size of *Sophophora* are positively correlated with transposable element abundance, whereas in *Drosophila* subgenus satDNA sequences strongly correlate with genome sizevariation.

Keywords: Satellite DNAs; Drosophila; Genome Evolution

Lista de Figuras

chromosome specific arrays (yellow) when compared to autosomal arrays (green). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

Figura 7 do Capítulo 2: A-B FISH with *CDSTR*130 (green) and pBuM (red) probes onto extended DNA fibers of *D. mojavensis*. (C) Schematic representation of *CDSTR130* and *pBuM* organization found on *contigs Ctg01_2999(*AAPU01002998.1) and *Ctg01_4375* (AAPU01004374.1 retrieved from the *D. mojavensis* assembled genome.

Figura 8 do Capítulo 2 : Transcription profile of satDNA families in *D. buzzatii* (A) and *D. mojavensis* (**B)** on five different developmental stages. Counts were normalized to one million reads.

Figura 9 do Capítulo 2: Representative ideogram showing the chromosomal localization of all satDNAs identified in *D. buzzatii*, *D. seriema* and *D. mojavensis*.

Figura 1 do Capítulo 3: Proportion (in %) in the genomes of the 36 *Drosophila* species of overall repetitive DNA and satellite DNA amount. The intensity of green (higher) and red (lower) colors are proportional to the variation inside column. The species are presented according to the phylogenetic tree topology as proposed by Russo et al. 2013, and we have indicated the genome sizes of each sequenced genome

Figura 2 do Capítulo 3: Genome size evolution and repeat composition of 36 *Drosophila* species and one subspecies of *D. mojavensis*. Repetitive DNA (red) and Satellite DNA (blue) proportions in Mb of *Drosophila* species calculated according to each species genome size (green).

Figura 3 do Capítulo 3: A. Monomer length of the 172 satellite DNA sequences described in the 36 species of *Drosophila* separated on 10 bp intervals. **B.** GC content variation of the monomers identified on 10% intervals. **C.** Overall distribution of satDNA families genomic proportion analyzed independently.

Figura 4 do Capítulo 3: A-C. Variation in satDNA library profile among close related species from: **(A)** *D. repleta* group; **(B)** *D. virilis* group; **(C)** *D. pseudoobscura* group. **D**. Variation in *1.688* satDNA abundance on 13 species of *D. melanogaster* group.

Figura 5 do Capítulo 3: Representative sequence alignment of *1.688* consensus of 13 *Drosophila* species described in this study indicating the presence of conserved regions. Dark blue indicates regions with higher sequence conservation while white regions indicate no conservation among the sequences.

Figura 6 do Capítulo 3: Correlation of repeats with genome size. Graphs show the Spearman's rank correlation between Repetitive Content and Genome Size; Repetitive Content

and SatDNA Content; and SatDNA Content and Genome Size in: **(a)** *Drosophila* genus, **(b)** Sophophora subgenus and **(c)** *Drosophila* subgenus.

Lista de Tabelas:

Sumário

1. INTRODUÇÃO

1.1. Paradoxo do valor C e DNAs repetitivos

Estudos realizados durante a primeira metade do século XX mostraram que a quantidade de DNA poderia diferir entre espécies e principalmente entre eucariotos e procariotos. Estes estudos demonstraram que algumas espécies de anfíbios e peixes continham cerca de vinte vezes mais material genético por núcleo haplóide (Valor C) do que espécies de mamíferos, dentre elas o homem (Gregory 2005). Esta aparente incongruência entre o grau de complexidade do organismo e o tamanho do genoma ficou conhecida na literatura como “Paradoxo do valor C” (Doolittle e Sapienza 1980; Moran e Morrish 2005). Um exemplo clássico deste paradoxo é o protozoário *Amoeba dubia*, que contém aproximadamente 200 vezes mais material genético do que a quantidade de material genético presente em uma célula humana.

O Paradoxo do valor C perdurou por alguns anos e começou a ser elucidado mesmo antes do advento das técnicas de sequenciamento e análise de genomas completamente sequenciados. Warring e Britten (1966), através do estudo da cinética de reassociação do DNA, demonstraram que a taxa de renaturação dos fragmentos de DNA depende do grau de similaridade entre as fitas da dupla-hélice. Assim, sequências de cópia única possuem taxas de reassociação mais lentas em relação a sequências menos complexas e/ou repetitivas. A identificação das sequências altamente repetitivas contribuiu para a inferência de que estas perfazem uma porção muito maior do genoma do que os genes codificantes de proteínas (Gregory 2005).

1.2. DNAs Satélites: características gerais

A estrutura e organização dos genomas de eucariotos são estudadas há várias décadas, e hoje é bem estabelecido que apenas uma pequena fração dos genomas é composta por sequências codificantes de proteínas. Diversos estudos ao longo das últimas

quatro décadas mostraram que a porção de DNA repetitivo é mais representativa do que os genes codificantes de proteínas (Gregory 2005; revisado por Richard e cols. 2008).

Elementos repetitivos estão amplamente distribuídos no genoma dos eucariotos e compõem, na maioria dos casos, cerca de 50% destes genomas. É possível classificar os elementos repetitivos em duas principais categorias: sequências repetidas em tandem e sequências repetidas dispersas, cada uma delas com diversas subdivisões. Entre os DNAs repetitivos dispersos destacam-se os elementos transponíveis, as duplicações segmentares, famílias gênicas e pseudogenes. Os DNAs repetidos em tandem são compostos basicamente por DNAs satélites, DNAs minissatélites, DNAs microssatélites e algumas famílias gênicas (Charlesworth e cols.1994; Richard e cols. 2008).

DNAs satélites são componentes característicos da grande maioria das espécies de eucariotos. Entre espécies, DNAs satélites são altamente variáveis em sequência de nucleotídeos, complexidade, tamanho da unidade de repetição e abundância, possuindo apenas duas características compartilhadas entre todos: a organização das repetições em longos arranjos em tandem e a localização predominantemente heterocromática (Plohl 2012).

Quando milhares de cópias homólogas (monômeros) são encontradas arranjadas em tandem ao longo de diferentes posições genômicas, estas cópias são consideradas uma “família” de DNA satélite. Esta definição também pode ser utilizada para DNAs satélites homólogos presentes em diferentes espécies. Vários DNAs satélites não-homólogos (diferentes famílias) podem ser encontrados no genoma de uma espécie (Lohe e Roberts 1993, Pezer e cols 2012). Um único DNA satélite, ou vários DNAs satélites somados, podem representar mais do que 30% do genoma de espécies de eucariotos (Bachmann e cols. 1996; Pons e cols. 2004).

Cópias de uma mesma família de DNA satélite podem ser encontradas em apenas um cromossomo (De la Herrán e cols. 2001), em vários (Bachmann e cols. 1989; Kuhn e cols.

2008), ou em todos os cromossomos de uma espécie (Mravinac e cols. 2004, Tyler-Smith e Willard1993). Mais de uma cadeia contendo cópias de uma mesma família de DNA satélite pode estar presente em um único cromossomo e mais de uma família de DNA satélite pode estar presente no mesmo cromossomo (Lohe e cols. 1993; Shiels e cols. 1997; Kuhn e cols. 2009). Por exemplo, em *D. melanogaster*, a região centromérica do cromossomo X é composta por cadeias contendo três DNAs satélites diferentes (Sun e cols. 2000).

Assim como as porções heterocromáticas do genoma, DNAs satélites são mais frequentemente encontrados em regiões centroméricas ou pericentroméricas (Mravinac e cols. 2004; Kuhn e cols. 2008), menos frequentemente em regiões subteloméricas (Ugarkovic e Plohl 2002; Heslop-Harrison e cols. 2003) e, mais raramente, em regiões intercalares (Nagaki e cols. 1999; Heslop-Harrison e cols. 2003; Kuhn e cols. 2012). Há ainda casos em que algumas poucas repetições foram encontradas em regiões eucromáticas (Pezer e cols. 2011; Kuhn e cols. 2012) (Figura 1).

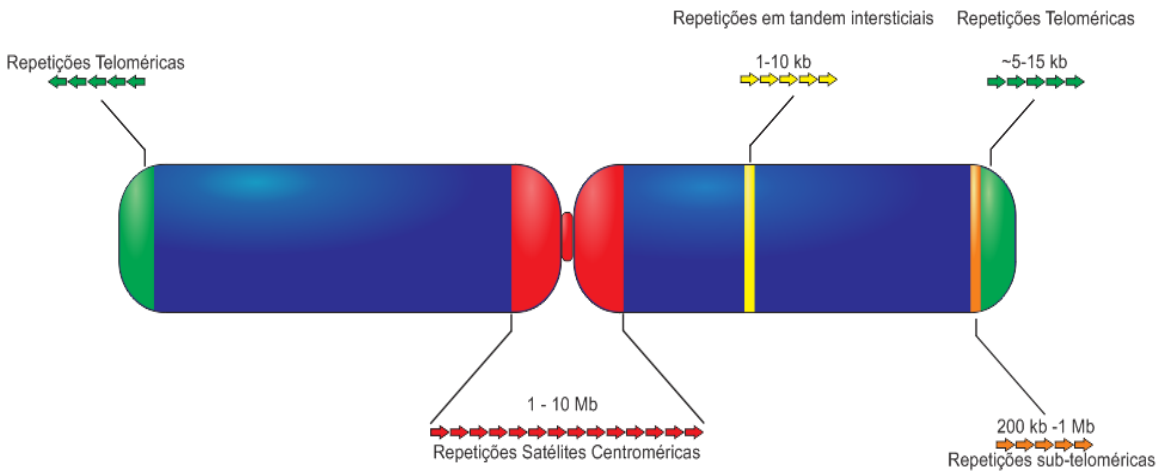

Figura 1. Representação esquemática de um cromossomo eucariótico e as sequências altamente repetidas em tandem (DNAs satélites) que normalmente compõem o genoma de uma espécie. São ilustradas as repetições satélites em tandem centroméricas, intersticiais ou eucromáticas, sub-teloméricas e teloméricas.

Outra característica bastante variável dos DNAs satélites é o tamanho das unidades de repetições que compõem os arranjos. Foram descritos DNAs satélites com unidades de repetição curtas de apenas 2 pb nos caranguejos *Cancer irroratus* e *C. borealis*, de 5 e 10 pb em *Drosophila melanogaster* ou de 5 pb em humanos (Gray e Skiner 1974; Lohe e Roberts 1988). Por outro lado, foram relatados DNAs satélites compostos por monômeros com tamanho superior a 1 kb, como é o caso do DNA satélite de cetáceos (Árnason e Grétarsdóttir 1992), com cópias de 1.740 pb. O caso mais extremo já caracterizado é do DNA satélite presente no cromossomo 8 humano que possui um monômero de 12 kb organizado em arranjos de até 1,7 Mb (Warburton e cols. 2008). Apesar desta enorme variação no tamanho de repetições, a maioria dos DNAs satélites descritos possui monômeros com sequências que variam entre 100 e 400 pb.

O número de cópias de uma família de DNA satélite também é bastante variável, podendo conter 10.000 cópias em *Glycine max* (Morgante e cols. 1997) ou até 6.000.000 cópias no roedor *Ctenomys haigi* (Slamovits e cols. 2001). Eventos de amplificação e deleção de sequências de DNAs satélites ocorrem com frequência, resultando em variação no número de cópias, como observado em espécies de roedores do gênero *Ctenomys* (Slamovits e cols. 2001). A variação também pode ocorrer entre indivíduos da mesma espécie, como é o caso do DNA satélite HSat3 de humanos, em que foi descrito uma variação de até 98 Mb no número de cópias do cromossomo Y entre indivíduos (Altemose e cols. 2014).

1.3. Evolução de DNAs satélites

DNAs satélites são compostos por longas cadeias de cópias em tandem não-codificantes de proteínas que estão localizadas predominantemente em regiões heterocromáticas. De acordo com a teoria de evolução molecular neutra, é esperado que sequências de DNA funcionais, como genes codificadores de proteínas, possuam uma

variabilidade menor quando comparadas a sequências com pouco ou nenhum papel funcional (Nei 1987). Desta forma, o processo evolutivo esperado seria o acúmulo aleatório de mutações ao decorrer dos arranjos de DNA satélites, gerando assim uma grande variabilidade nucleotídica dentro das cadeias.

No entanto, as análises de sequências de inúmeros DNA satélites apresentam um padrão oposto ao esperado por evolução molecular neutra, no qual é observado uma alta taxa de similaridade entre cópias de um mesmo DNA satélite presentes em uma espécie (revisado em Plohl 2012). Por outro lado, existe uma grande divergência interespecífica entre cópias de uma mesma família de DNA satélite. Isto significa que existe algum mecanismo molecular que leva as cadeias de DNA satélite à homogeneização, ou seja, serem extremamente similares entre si dentro de uma dada espécie. No caso de isolamento reprodutivo, este processo pode levar à homogeneização de mutações espécie-específicas, resultando em uma variabilidade intraespecífica menor do que a observada entre cópias de duas ou mais espécies, um fenômeno descrito na literatura como “evolução combinada” ou “evolução em concerto” (Zimmer e cols. 1980) (Figura 2).

O fenômeno de evolução combinada também é observado em diferentes famílias de DNAs repetidos em tandem, como famílias gênicas de histonas, globinas, imunoglobulinas, além de diversas sequências não-codificantes (Liebhaber 1981; Scott 1984; Kuhn e cols. 2008; 2012). O baixo grau de variabilidade entre cópias é explicado pela ação de vários mecanismos moleculares, como a recombinação desigual, a conversão gênica, a derrapagem da polimerase (*replication slippage*) e a amplificação via “círculo-rolante” de DNA extracromossômico circular (Smith 1976; Dover 1986, Cohen e Segal 2009). A influência destes mecanismos na evolução molecular é conhecida como impulso molecular ou *molecular drive* (Dover 1982).

Algumas evidências sugerem que o processo de homogeneização é mais eficiente entre cópias presentes na mesma cadeia, ou cromossomo, do que entre cópias presentes em diferentes cromossomos (Willard e Waye 1987; Hall e cols. 2005; Rudd e cols. 2006).

Este fenômeno pode ser explicado pela maior probabilidade de ocorrência de recombinação e, consequentemente, de homogeneização entre cópias de uma mesma cadeia do que entre cópias de cadeias em diferentes cromossomos (Dover 1982). Este tipo de situação pode levar ao surgimento de subfamílias de DNAs satélites cromossomo-específicas, como observado no DNA satélite alfóide humano (Willard e Waye 1987; Rudd e cols. 2006).

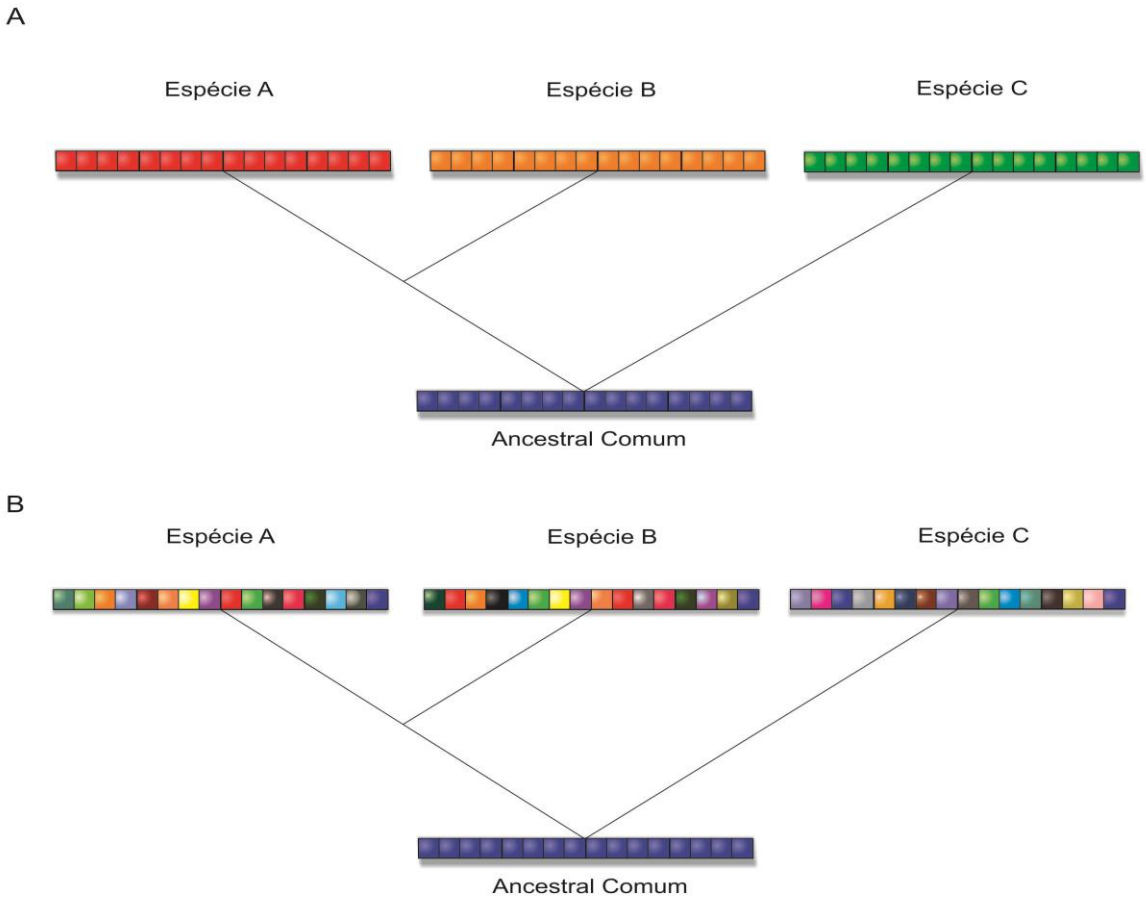

Figura 2. (A) Evolução combinada através da homogeneização de mutações espécie-específica em todas as cópias de um arranjo. (B)Padrão esperado de evolução das repetições de acordo com a teoria clássica de evolução neutra.

Além dos mecanismos de homogeneização das sequências, há os mecanismos responsáveis pela dispersão de sequências de DNAs satélites por todo o genoma. Sequências de alguns elementos transponíveis apresentam alta similaridade com

sequências de DNAs satélites, indicando que elementos móveis podem ser a fonte de origem e/ou dispersão para alguns DNAs satélites, como foi demonstrado em *Drosophila obscura* (Miller e cols. 2000).

Outro importante mecanismo de amplificação e dispersão de DNAs satélites ocorre através de DNAs extracromossômicos circulares (eccDNA). Estes eccDNAs são formados através de eventos de recombinação ectópica entre cópias de uma mesma cromátide, o que acarreta uma deleção de cópias que passam a constituir um DNA circular. Estas moléculas de DNA circular contendo cópias de DNA satélite podem ser amplificadas, caso haja uma região iniciadora de replicação dentro da molécula, e posteriormente inseridas no mesmo ou em outro local do genoma (Cohen e Segal 2009).

1.4. Papel Biológico dos DNAs satélites

Estudos realizados nas décadas de 70 e 80, e confirmados até o presente, revelaram que os DNAs satélites não codificam proteínas e que se concentram em regiões caracterizadas por baixos níveis de recombinação e alta compactação da cromatina. Com base nestes dados, DNAs satélites foram historicamente rotulados como “DNA entulho” e sua presença no genoma foi explicada por sua habilidade de amplificação (Orgel e Crick 1980).

Embora os monômeros de DNAs satélites apresentem uma baixa variabilidade entre si, algumas regiões dos monômeros apresentam uma maior conservação de sequência, enquanto outras possuem uma maior variabilidade (Mravinac e cols. 2004). Esta distribuição não uniforme da variabilidade na sequência da unidade de repetição indica a presença de pressões seletivas para a conservação de determinadas regiões possivelmente funcionais (Heslop-Harrison e cols. 1999; Pezer e cols 2011). Pequenos motivos conservados foram descritos em DNAs satélites centroméricos de arroz (*Oryza sativa*) e milho (*Zea mays*),

sugerindo a manutenção de um mesmo segmento funcional entre espécies separadas há pelo menos 50 milhões de anos (Cheng e cols. 2002).

O conceito de função biológica é mais claramente compreendido quando são analisadas sequências de DNA altamente transcritas ou que estejam envolvidas em processos bioquímicos fundamentais para o organismo (Graur e cols. 2015). Apesar de DNAs satélites não codificarem proteínas, é crescente o número de estudos que indicam que DNAs satélites podem estar envolvidos em diversas funções biológicas, como na formação do centrômero (Henikoff e cols. 2001), modulação da cromatina (Volpe e cols. 2002), e regulação gênica/desenvolvimento (Ugarkovic 2005). Sítios conservados de 17 pb denominados “CENP-B box” estão presentes nos monômeros de 171 pb do DNA satélite α de humanos, que, por sua vez, está localizado no centrômero de todos os cromossomos. Os sítios de 17 pb que compõem o “CENP-B box” facilitam a interação entre as proteínas formadoras do complexo cinetocórico e a proteína centromérica B (Masumoto 1989).

A transcrição de DNAs satélites já foi descrita em diversos invertebrados, vertebrados e plantas (Araújo e cols. 2017; De Lima e cols. 2017). A transcrição geralmente ocorre durante estágios iniciais do desenvolvimento e ainda pode ocorrer diferencialmente entre tecidos diferentes, sugerindo um papel regulatório (revisado por Ugarkovic 2005). Alguns trabalhos também sugerem que transcritos de DNA satélite atuam em mecanismos transregulatórios e que os monômeros possuem potenciais regiões promotoras da transcrição para as enzimas RNA polimerase II e III (Renault e cols. 1999; Pezer e cols. 2011).

Recentemente, foi demonstrado que os transcritos do DNA satélite mais abundante de *D. melanogaster*, chamado de *1.688*, desempenham importantes funções biológicas. Rosic e cols. (2014) demonstraram que transcritos do DNA satélite *1.688* ligam-se à proteína centromérica CENP-C, a qual tem um papel essencial na formação e funcionalidade centromérica nesta espécie. Além da interação com as proteínas centroméricas, Menon e

cols. (2014) mostraram que transcritos do mesmo DNA satélite contribuem para a regulação de compensação de dose no cromossomo X de machos desta espécie. Foi observado que os transcritos do *1.688* ancoram em sítios promotores do complexo regulador MSL (*male specific lethal*), resultando em uma duplicação na expressão dos genes presentes no cromossomo X dos machos.

1.5. O Gênero *Drosophila*

A família *Drosophilidae* é composta por mais de 3.600 espécies, sendo mais de 1.400 delas pertencentes ao gênero *Drosophila* (Powell 1997; O'Grady e Markow 2009). Espécies do gênero *Drosophila* estão distribuídas por praticamente todo o globo terrestre, em regiões temperadas, tropicais e desertos. O gênero *Drosophila* é parafilético, já que vários outros gêneros estão incluídos na filogenia de *Drosophila* (Russo e cols. 2013). O gênero *Drosophila* é subdividido em 15 subgêneros, dos quais destacam-se os subgêneros *Drosophila* e *Sophophora* pela alta diversidade de espécies (Powell 1997). De acordo com dados biogeográficos e fósseis, a separação entre os dois subgêneros ocorreu entre 40 e 60 milhões de anos atrás (Powell e DeSalle1995; Powell 1997). Estes subgêneros ainda são subdivididos taxonomicamente em radiações e grupos de espécies.

Dos dois subgêneros principais, *Drosophila* é o maior. Tem uma ampla distribuição e alguns de seus membros mostram nichos ecológicos interessantes, como fungos e cactos (Markow e O'Grady 2008). Este subgênero é composto por três linhagens principais: 1) o grupo de espécies *funebris*; 2) a radiação *virilis-repleta*; e 3) a radiação *immigrans-tripunctata* (Val e cols. 1981). A posição filogenética do grupo funebris não está bem resolvida e alguns estudos sugerem que faz parte da radiação *immigrans-tripunctata* (Russo e cols. 2013). Deve-se ressaltar que o subgênero *Drosophila* é parafilético e inclui os *Drosophilidae* havaianos (*Drosophila* havaianas + gênero *Scaptomyza*), que formam um grupo monofilético irmão da radiação *virilis-repleta* (Van der Linde e cols. 2010). A linhagem de *Drosophila* havaianas compreende aproximadamente 1000 espécies e forma uma própria

radiação adaptativa, com uma grande diversidade de formas e nichos ecológicos (Craddock e cols. 2016).

Por outro lado o subgênero Sophophora contém oito grupos de espécies dos quais destacam-se os grupos *obscura, saltans, willistoni, ananassae, montium* e *melanogaster* (Da Lage e cols. 2007). O subgênero Sophophora tem sido tradicionalmente dividido em dois, relacionando as espécies que possuem origem de diversificação no "velho mundo" como os grupos *obscura, ananassae, montium* e *melanogaster*, enquanto os grupos *saltans* e *willistoni* são considerados monofiléticos e formam um clado associado à colonização do continente americano (Russo e cols. 2013).

1.6 Genomas sequenciados e o estudo de DNAs satélites em Drosophila

A espécie *D. melanogaster* é um dos modelos biológicos mais importantes na história da ciência. Portanto, como era de se esperar, o genoma de *D. melanogaster* estava entre os primeiros a serem sequenciados entre os metazoários (atrás de *Caenorhabditis elegans*), e foi o primeiro genoma animal a ser sequenciado usando a abordagem de *shotgun* (Adams e cols. 2000). Tal importância experimental fomentou o sequenciamento de diversas espécies próximas a *D. melanogaster*.

O gênero *Drosophila* como um todo possui a maior quantidade de espécies com genomas sequenciados e disponíveis publicamente para estudo. Este fator fornece um material valioso para biólogos evolutivos, permitindo a caracterização de vários elementos genéticos e a determinação de sua organização, função e dinâmica evolutiva (*Drosophila* 12 Genomes Consortium 2007). Entre as 24 espécies com genomas sequenciados e montados (até julho de 2017), há pelo menos uma representante de cada grande grupo taxonômico de *Drosophila*.

Apesar do grande repertório de genomas sequenciados disponíveis para análise entre espécies de *Drosophila*, até o presente apenas 23 famílias de DNAs satélites de *Drosophila* estão depositadas nos bancos de dados Genbank ou Repbase (acessados em maio de 2017).

Em *Drosophila*, as investigações de DNAs satélites começaram relativamente mais tarde do que outras estatísticas de animais, porém obtiveram um avanço significativo em um curto período de tempo (Laird e McCarthy 1968, Gall e cols., 1971). Os estudos iniciais de DNAs satélites em *Drosophila* usavam o gradiente de densidade de CsCl para caracterizar as sequências mais abundantes em genomas. As primeiras espécies que obtiveram uma caracterização prévia foram *D. virilis*, *D. melanogaster* e *D. hydei* (Gall e cols. 1974; Barnes e cols. 1978; Renkawitz 1979). Nas últimas décadas, no entanto, DNAs satélites foram estudados principalmente a partir de uma pequena amostra de repetições clonadas obtidas por abordagens experimentais com viés de análises (geralmente por digestão de restrição e / ou PCR) isoladas de uma ou poucas espécies. (Brutlag e cols. 1977; Waring e Pollack 1987; Bonaccorsi e Lohe 1991; Bachmann e Sperlich 1993; Kuhn e cols. 1999; Kuhn e cols. 2008). A maioria dos estudos publicados recentemente centrou-se na descrição da dinâmica evolutiva ou influência de famílias de DNAs satélites previamente descritos nos subgrupos *D. melanogaster* e *D. virilis* (Kuhn e cols. 2012; Garavis e cols. 2015; Dias e cols. 2014; Gallach 2014; Larracuente 2014; Khost e cols. 2017; De Lima e cols. em preparação).

Em *Drosophila* os DNAs satélites podem representar mais de 30% do genoma e os eventos de amplificação/contração de famílias distintas de DNA satélite foram identificados como um fator importante na arquitetura e do tamanho do genoma de *Drosophila* (Bosco e cols. 2007). Além disso, o processo de especiação pode estar associada à evolução de DNA satélite dado que mudanças rápidas no número de cópias podem desencadear a rápida evolução do genoma e gerar barreiras reprodutivas, como observado entre *D. melanogaster* e *D. simulans* (Ferree e Barbash 2009).

Apesar do esforço em descrever as sequências de DNA satélite em *Drosophila*, os estudos focados em questões evolutivas específicas são escassos e realizados em um número de espécies e grupos de espécies pouco representativos (revisado por Plohl 2012). Como resultado, a maioria dos genomas recentemente sequenciados, se não todos, carecem de uma caracterização das sequências DNA satélite e usam o banco de dados *RepeatMaske*r para identificar a presença geral de famílias repetitivas. Tendo em vista a subrepresentação de sequências de DNA satélites nos bancos de dados, os dados de montagem e anotação genômica tendem a negligenciar uma grande fração do genoma destas espécies.

Atualmente, muitas questões sobre DNAs satélites continuam sem respostas, embora muito tenha sido teorizado sobre a sua origem, função e evolução e avanços significativos tenham sido feitos no decorrer das últimas quatro décadas (Plohl e cols. 2012; revisado em Garrido-Ramos 2015). Apesar de sua importância para a organização, função e evolução do genoma, os DNAs satélites raramente foram estudados em espécies de *Drosophila* com genomas sequenciados e em eucariotos em geral. Com o crescente número de evidências do papel biológicos das sequências de DNAs satélites, se faz necessária à descrição e caracterização das diferentes sequências repetitivas presentes em um genoma. Ademais, a aplicação de abordagens genômicas têm sido fundamental para avanços importantes no conhecimento do papel dos DNAs satélites na evolução dos genomas eucarióticos (Melters e cols. 2012; Zhang e cols. 2014; Feliciello e cols. 2015). Logo, a identificação de DNAs satélites em uma escala filogenética em *Drosophila* possibilita a ampliação e o refinamento na caracterização e análise destes elementos genéticos.

Deste modo, a análise dos genomas sequenciados de espécies do gênero *Drosophila* fornece um material valioso para estudos de genômica comparativa, permitindo uma caracterização ampla de elementos genéticos historicamente negligenciados, além da

determinação da organização, função e dinâmica evolutiva de um dos principais componentes do genoma de eucariotos.

2. OBJETIVOS

2.1. Objetivo Geral

Neste trabalho combinamos ferramentas de bioinformática, biologia molecular e de citogenética molecular para um estudo integrativo e aprofundado dos DNAs satélites presentes em *Drosophila* com genomas sequenciados.

De posse destes dados, tivemos como objeto ampliar de forma significativa o entendimento sobre a organização, a dinâmica evolutiva e papel biológico dos DNA satélites e o impacto deste componente genômico na arquitetura e diferenciação dos genomas eucarióticos.

2.2. Objetivos específicos

- Identificação e caracterização dos DNAs satélites de 36 espécies de Drosophila que possuem dados genômicos disponíveis em bancos de dados públicos.

- Estudar a estrutura, variação e evolução dos DNAs satélites de cada uma das espécies e comparar os padrões encontrados com modelos existentes de evolução de DNAs satélites;

- Estimar a contribuição genômica para cada DNA satélite identificado e correlacionar com o percentual total de DNAs repetitivos;

- Investigar a presença (ou não) de regiões nucleotídicas conservadas dentro das cópias de DNA satélites e a presença (ou não) de regiões promotoras;

- Determinar a localização e organização genômica dos DNAs satélites identificados em cromossomos metafásicos e politênicos nas espécies *D. mojavensis*, *D. buzzatii* e *D. seriema*;

-Determinar se a variação de DNAs satélites nos genomas de 36 espécies de *Drosophila* tem correlação com o processo de variação no tamanho do genoma

3. MATERIAL E MÉTODOS

3.1. Dados genômicos

Para a busca por sequências de DNAs satélites presentes no genoma das espécies de *Drosophila* foram utilizados quatro tipos de dados genômicos: (i) dados de contigs "congelados" de 454 Roche platform (ii) *reads* de *Next Generation Sequencing* - NGS(iii) *reads* de *Pacific Bioscience Sequencing* e (iv) arquivos de genomas montados presentes nos bancos públicos *GenBank* e *Flybase.org*.

O genoma montado de *D. buzzatii* foi disponibilizado para nosso grupo em maio de 2013 pelo Prof. Alfredo Ruiz (Universidade Autonoma de Barcelone). Foram obtidas as sequências dos contigs montados utilizando *reads* 454 e, posteriormente, foram usadas as reads de *pair-end* de *Illumina HiSeq* 200 (7x e 76x de cobertura, respectivamente)(Guillen e cols. 2015).

Os dados genômicos de sequenciamento de *D. mojavenis* foram obtidos utilizando a plataforma *Illumina HiSeq* 200 (20x de cobertura). Estes dados foram gentilmente cedidos pelo Prof. Bernardo de Carvalho do Laboratório de Genética de Populações de Drosófila da UFRJ. O último lançamento público do genoma completo de *D. mojavensis* está disponível para análises nos bancos de dados de livre acesso http://flybase.org/.

Para a espécie *D. seriema* foi realizado um experimento de sequenciamento visando produzir os dados necessários para a análise dos DNAs repetitivos presentes no genoma desta espécie. Os dados genômicos foram obtidos com a plataforma *Illumina MiSeq* implementado no Departamento de Biologia Geral da UFMG. A biblioteca genômica foi gerada utilizado um *pool* de indivíduos machos e fêmeas e com *reads pair-end* de ~300 pb. O resultado do sequenciamento gerou um arquivo de sequências referentes a 25x de cobertura genômica desta espécie.

As demais sequências utilizadas neste estudo foram retiradas do banco de dados Short Reads Archive (SRA) do NCBI (https://www.ncbi.nlm.nih.gov/sra). A lista completa das sequências utilizadas e os respectivos números de acesso estão disponíveis na Tabela 1.

Tabela 1. Lista das 36 espécies do gênero *Drosophila* analisadas neste trabalho juntamente com os respectivos dados do número de reads utilizadas, cobertura do genoma e arquivos de NGS utilizados.

Espécie	Subgênero	Total reads analisadas	Cobertura do Genoma	Arquivos NCBI/SRA
D. affinis	*Sophophora*	3534015	1,7670	ERX103525
D. albomicans	*Drosophila*	2671499	1,0685	SRX010928
D. americana	*Drosophila*	1326207	0,4736	SRX2582810
D. ananassae	*Sophophora*	5072230	2,5878	SRX144727
D. arizonae	*Drosophila*	5027345	3,5403	SRX1065941
D. biarmipes	*Sophophora*	6693021	3,3465	SRX095626
D. bipectinata	*Sophophora*	1666248	0,8167	SRX094522
D. burlai	*Sophophora*	1823514	-	SRX883275
D. busckii	*Dorsilopha*	9243029	6,6021	SRX265057
D. buzzatii	*Drosophila*	2165197	1,2811	N/A
D. elegans	*Sophophora*	2408863	1,2546	SRX094536
D. erecta	*Sophophora*	5045489	3,1732	SRX997779
D. eugracilis	*Sophophora*	2392509	1,0493	SRX095624
D. ficusphila	*Sophophora*	4884243	2,5571	SRX095448
D. hydei	*Sophophora*	1000000	0,4835	DRX055253
D. kikkawai	*Sophophora*	1801891	0,85804	SRX095451
D. leontia	*Sophophora*	2227008	-	SRX883299
D. malerkotliana	*Sophophora*	1431137	0,7015	SRX237739
D. mauritiana	*Sophophora*	2243821	1,4201	SRX183513
D. melanogaster	*Sophophora*	4974307	2,8587	SRX1961048
D. mojavensis wrigley	*Drosophila*	2174346	1,2865	SRX2932915
D. mojavensis baja	*Drosophila*	3158945	1,8691	SRX1065937
D. montana	*Drosophila*	1572497	-	SRX1604922
D. novamexicana	*Drosophila*	1000000	0,4	SRX2582808
D. orena	*Sophophora*	3959570	1,4141	SRX997798
D. pseudoobscura	*Sophophora*	1519783	0,7874	SRX204754
D. persimilis	*Sophophora*	1000000	0,51817	SRR363439
D. rhopaloa	*Sophophora*	1628575	0,8438	SRX095455
D. santomea	*Sophophora*	3189067	1,8649	SRX752500
D. sechellia	*Sophophora*	6918830	4,1679	SRX287396
D. seriema	*Drosophila*	2144275	-	ERX2037878
D. simulans	*Sophophora*	1260769	0,7899	SRX1799314
D. subobscura	*Sophophora*	6337592	4,2250	DRX055266
D. takahashii	*Sophophora*	2049860	0,9902	SRX095457
D. teissieri	*Sophophora*	3929041	2,36689	SRX854063
D.virilis	*Drosophila*	1000000	0,3076	SRX669289
D. yakuba	*Sophophora*	3829418	2,2525	SRX494771

3.2. Identificação e mineração das sequências de DNAs satélites

3.2.1. Identificação de sequências altamente repetidas em tandem utilizando scripts em *BioPerl*.

A identificação de sequências repetitivas foi dividida em duas etapas sendo a primeira delas restrita à espécie *D. buzzatii*. Nesta etapa foram utilizados 12 *scripts* em *Perl* sendo 11 desenvolvidos neste estudo para a identificação de sequências altamente repetidas em tandem presente em genomas montados. Esta etapa foi desenvolvida em parceria com o Prof. Jerônimo Ruiz no Laboratório de Informática de Biossistemas no Centro de Pesquisas René Rachou - Fiocruz - Minas Gerais em Belo Horizonte. Os *scripts* foram idealizados com a finalidade de isolar e caracterizar sequências altamente repetidas em tandem.

A primeira etapa de identificação das repetições em tandem em 49 mil contigs utilizou o software Tandem Repeat Finder 4.04 (Benson 1999) em linha de comando com os parâmetros de 1,1,2,80,5,200 e 750 para *match*, *mismatch*, *indel*, probabilidade de *match*, probabilidade de *indel*, *score* mínimo e *score* máximo, respectivamente. Repetições com tamanho inferior a 50 pb foram eliminadas da análise devido à falta de confiabilidade gerada nas etapas seguintes, retirando assim dados que podiam induzir a análises errôneas. A etapa seguinte visou eliminar a redundância gerada pelo TRF através de várias etapas de clusterização. Brevemente, foram realizadas quatro rodadas de clusterização em que foram agrupadas todas as sequências com pelo menos 70% de identidade e com mais de 60% de cobertura após uma análise de *blast2seq all-vs-all*. Uma etapa de clusterização por alinhamento também foi realizada utilizando CAP3. Todas as sequências consensos derivadas deste processo foram usadas como *query* em busca por similaridade contra os contigs resultantes da montagem final do genoma de *D. buzzatii*. O *output* indicou as sequências que obtiveram o maior número de *hits* em ordem de abundância. Posteriormente, os possíveis DNAs repetitivos foram classificados de acordo com as respectivas abundâncias, proporção representativa do genoma e conteúdo GC%. O

conjunto de todos os scripts gerou um *pipeline* utilizado para a análise (Anexo 1). A última parte desta metodologia envolveu uma verificação manual de todos os possíveis DNAs repetitivos em tandem nos bancos de dados *RepBase* (Jurka e cols. 2005) com o propósito de eliminar possíveis *hits* pertencentes a TEs ou duplicações segmentares de sequências não satélites.

3.2.2. Identificação de sequências de DNA satélite utilizando o pipeline *RepeatExplorer*

Para identificação de sequências repetitivas presentes nos genomas de *Drosophila* foi utilizada a plataforma de *Repeat Explorer* (Novak e cols. 2013). Esta plataforma é uma coleção de ferramentas de caracterização de elementos repetitivos agrupadas em um *pipeline,* no qual um algoritmo de clusterização, baseado em gráficos de similaridade, utiliza sequências de *reads* curtas para gerar clusters contendo sequências repetitivas similares entre si. Desse modo, a identificação *de novo* de sequências de DNA satélites pode ser realizada sem a necessidade de bancos de dados de referência e com baixa cobertura genômica (Novak e cols 2010).

Os dados utilizados para a identificação *de novo* dos DNAs satélites presentes em *Drosophila* foram baixados diretamente do EBI Short Reads Archive usando a ferramenta *Get Data –> EBI SRA tool*. É importante ressaltar que estes arquivos são compartilhados com o banco de dados SRA do NCBI. Os dados presentes em ambos os bancos de dados estão depositados no formato FASTQ, o qual combina a sequência de nucleotídeos com a informação dos valores de qualidade do sequenciamento para cada nucleotídeo. Posteriomente ao *upload* dos dados brutos, a primeira etapa da análise é o adequamento de todos os arquivos para o mesmo formato de qualidade *FASTQ-Sanger*, realizado pela ferramenta *FASTQ Groomer*. Na segunda etapa, todas as reads foram padronizadas com 100 pb e todos os adaptadores de sequenciamento encontrados nos dados depositados foram removidos para evitar viés de representação de sequências. Também foi realizado

um *cut-off* 30 de qualidade de reads, e somente reads com 90% ou mais de bases com valores superiores ao cut-off foram utilizadas nas análises.

Tendo em vista que as análises de clusterização requerem como input um único arquivo contendo as *reads* pré-processadas no formato FASTA, todos arquivos utilizados foram convertidos do formato FASTQ para FASTA pela ferramenta *FASTQ to FASTA converter*. Em seguida, foram adicionados prefixos específicos para cada espécie contendo as três primeiras letras do respectivo nome da espécie (p. ex: Dmel; Dvir) através da ferramenta *Repeat Explorer –> (UTILITIES) Rename Sequences*.

A última etapa consistiu no processo de clusterização das sequências atráves de uma comparação de similaridade entre todas as reads de cada espécie. A identificação das sequencias altamente abundantes se baseou na formação de clusters entre as reads que apresentaram uma maior similaridade de sequência entre si. Para que cada read fosse agrupada dentro de um mesmo cluster foi utilizado um limiar de 90% de similaridade de sequência entre pelo menos 65% de cada read. Após o processo de clusterização, todas as reads dos clusters são montadas utilizando o software CAP3 (Huang e Madan 1999) para gerar um arquivo contendo diversos contigs representativos de cada sequência identificada.

Posteriormente, todos os clusters foram analisados individualmente de acordo com o formato gerado pela divergência entre as sequências. Sequências altamente repetidas em tandem costumam fomar um padrão gráfico associado à densidade de sequências por cluster, isto é, quanto menor for a distância nucleotídidica observada entre as reads, mais denso será o *layout* do cluster. Sequências repetitivas em tandem apresentam caracteristicamente clusters com altos valores de densidade nucleotídica e gráficos igualmente condensados (Figura 3). Foram analisados todos os clusters com padrões gráficos condizentes com sequências repetitivas em que apresentaram um valor de composição genômica superior a 0,01%. Após a etapa de clusterização, todos os clusters foram identificados por buscas no banco de dados de sequências repetitivas RepBase ou pela identificação de domínios protéicos associados a elementos transponíveis.

No entanto, devido à baixa representatividade de DNAs satélites nos bancos de dados, as famílias de DNA repetidos em tandem foram confirmadas após uma análise manual utilizando o método de alinhamento comparativo entre duas sequências implementado no applet Dotlet (Junier e Pagni 2000). Estas análises foram utilizadas para a confirmação das sequências em tandem e posterior caracterização dos monômeros.

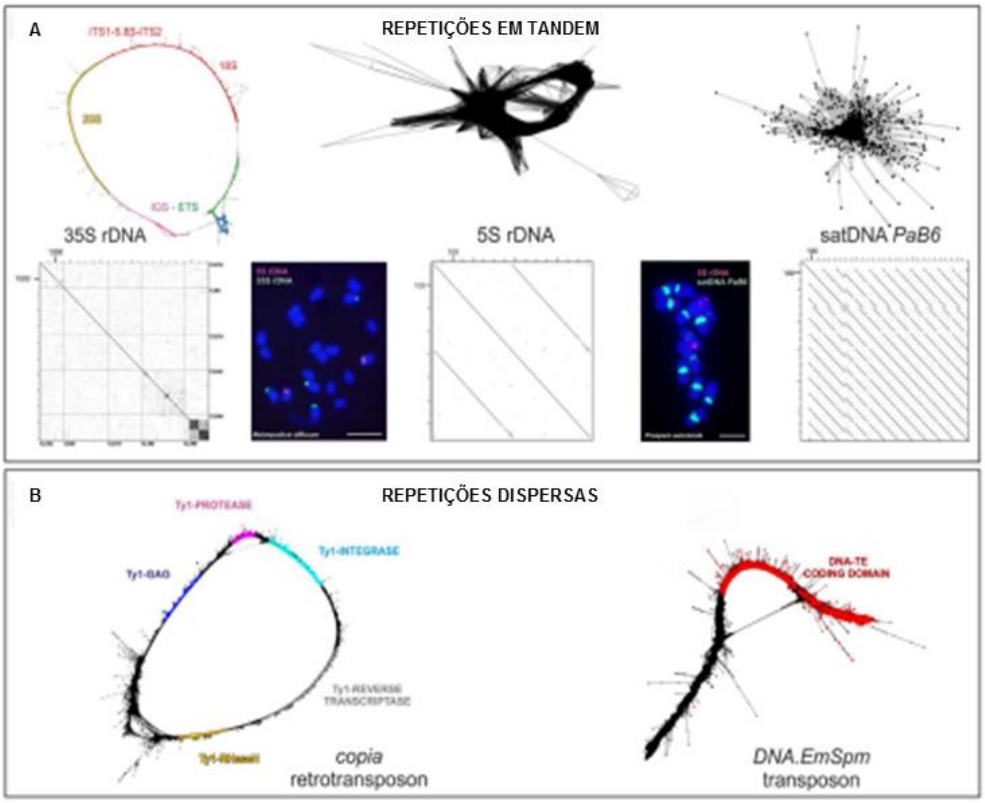

Figura 3. Exemplos de resultados baseados na distância nucleotídica após o processo de clusterização para cada tipo de sequências repetitivas identificadas pelo pipeline *RepeatExplorer*. **A.** Estruturação gráfica de sequências repetidas em tandem (rDNAs e DNAs satélites) e o padrão de organização de cada uma de acordo com a análise manual do *Dotplot*. **B.** Estruturação gráfica de sequências repetitivas dispersas (elementos transponíveis). Figura adaptada de Weiss-Scheneeweiss e cols. (2015).

3.3. Alinhamentos e análises de sequências do DNA satélite

As múltiplas cópias de DNAs satélite foram alinhadas com o auxílio do algoritmo do *Muscle* (Edgar 2004), presente no software MEGA 7 (Tamura e cols. 2016), com otimização feita manualmente quando necessária. O programa MEGA 7 foi utilizado também para análises do tamanho, composição nucleotídica e grau de variabilidade entre as cópias de

cada DNA satélite. O programa *Dotlet* foi utilizado para análise interna das cópias de DNA satélites, através da procura de repetições internas (diretas ou invertidas).

Árvores filogenéticas foram construídas pelo algoritmo *Neighbor Joining* (Saitou e Nei 1987) presente no programa MEGA 7(Tamura e cols. 2011). A avaliação estatística de cada ramo da árvore foi realizada através de análise de "*bootstrap*" (500-1000 réplicas dependendo da quantidade de sequências analisadas).

3.4. Análises Moleculares

3.4.1. Extração de DNA genômico

Para a extração do DNA de *D. mojavensis*, *D. buzzatii* e *D. seriema* foi utilizado o kit de extração Wizard® Genomic DNA Purification (Promega). A extração envolveu 30-50 moscas adultas, as quais foram previamente maceradas em nitrogênio líquido com micropistilos estéreis.

3.4.3. Eletroforese em géis de agarose

A concentração, pureza e tamanho dos DNAs analisados foram verificados em géis de agarose 1%-1,5% diluída em tampão TAE. As amostras de DNA foram preparadas pela adição de 5 µL do tampão de carreamento Loading Dye 6x (Promega). As eletroforeses foram realizadas a 70mA por 55 a 70 minutos e os géis foram corados com brometo de etídeo e visualizados e fotografados em transluminador UV.

3.4.4. Eluição de amplicons de PCR de géis de agarose

As eluições de fragmentos de DNA amplificados por PCR foram realizadas utilizando o *kit* "Wizard® SV Gel and PCR Clean-Up System" (Promega).

3.4.5. Ligação do DNA amplificado em plasmídios vetores

A ligação dos produtos obtidos através de PCR foi realizada com o *kit* "pGEM-T easy vector" (Promega).

3.4.6. Transformação bacteriana e sequenciamento dos clones

Os plasmídeos obtidos foram inseridos em células competentes (JM109-Promega) e os clones recombinantes foram selecionados para sequenciamento. O sequenciamento das

amostras foi terceirizado e realizado no Laboratório de Sequenciamento Mylleus em Belo Horizonte-MG. O aparelho utilizado para o sequenciamento das amostras foi um sequenciador ABI Prism 3130 (Life Technologies). Os cromatogramas fornecidos como resultados foram analisados no programa CHROMAS Lite 2.1.

3.5. Localização e organização genômica

Preparações de cromossomos metafásicos (a partir dos gânglios cerebrais) e politênicos (a partir de glândulas salivares), foram obtidas de larvas de *Drosophila* do 3º estágio. Fibras de DNA foram isoladas a partir de indivíduos adultos. As metodologias necessárias para as preparações cromossômicas e fibras de DNA de *Drosophila* foram descritas em Kuhn e cols. (2008). Sondas de DNAs satélite foram hibridadas nos cromossomos metafásicos, politênicos e fibras de DNA de *Drosophila* através da hibridização *in situ* fluorescente (FISH), de acordo com metodologia descrita em Kuhn e cols. (2008).

A obtenção de sondas foi realizada através da marcação do DNA inserido dentro de plasmídeos com biotina ou digoxigenina, de acordo com as instruções dos fabricantes. A hibridização *in-situ* e detecção dos sinais foram realizadas de acordo com metodologia descrita em Kuhn e cols. (2008). Para a detecção de sondas marcadas com biotina ou digoxigenina, foram utilizados fluorocromos como FITC ou rodamina ligados aos anticorpos anti-DIG/FITC e Neutravidina/Rodamina. Os cromossomos foram corados com DAPI (4',6-diamidino-2-phenylindole, dihydrochloride salt). As preparações foram analisadas em microscópio de epifluorescência Zeiss Axiophot 2 equipado com uma câmera CCD e as imagens foram obtidas com o programa AxioVision (Zeiss). Para a determinação do tamanho das cadeias de DNAs satélites, os tamanhos das fibras com sinais de hibridização foram medidos de acordo com protocolo descrito por Schwarzacher e Heslop-Harrison (2000).

3.6. Linhagens de *Drosophila*

Neste trabalho foram analisadas as mesmas linhagens utilizadas para os experimentos de sequenciamento de *Drosophila mojavensis* (linhagem: CI 12 IB-4 g8) , *D. buzzatii* (linhagem: St01) e *D. seriema* (linhagem: D73C3B). Estas linhagens são mantidas em nosso laboratório em frascos contendo meio de cultura a base de milho, farinha de soja e levedura de cerveja.

4. Capítulo I - Identificação das sequências repetitivas nos genomas de *D. buzzatii* com *pipeline* específico.

Identificação das sequências repetitivas nos genomas de *D. buzzatii* com *pipeline* específico.

A metodologia inicial de identificação de sequências repetitivas através de um *pipeline* específico foi utilizada somente para as sequências presentes em 52.747*contigs* utilizados para a montagem do genoma de *D. buzzatii*. Os *scripts* desenvolvidos demonstraram eficácia na identificação inicial de sequências repetitivas, embora tenham demonstrado a necessidade de curadoria manual de todos os dados, além das buscas por similaridades nos bancos de dados de DNAs repetitivos.

Como resultados foram desenvolvidos ou adaptados 11 diferentes scripts gerados na linguagem computacional *BioPerl* (Anexos 1 e 2).

A primeira etapa de identificação de sequências repetidas em tandem em *D. buzzatii* gerou uma biblioteca de 1.310 possíveis sequências repetitivas acima de 50 pb, das quais foram analisadas as 277 mais abundantes. Depois de análises visuais para eliminação de redundância e buscas por similaridade com TEs foram identificadas duas famílias de DNAs satélites altamente abundantes no genoma de *D. buzzatii*.

A família de DNA satélite mais abundante identificada em *D. buzzatii* apresentou um monômero de 189 pb. O monômero identificado corresponde cópias alfa do DNA satélite *pBuM*, previamente descrito em espécies do *cluster buzzatii*. Para identificar e caracterizar o *pBuM* no genoma montado de *D. buzzatii* foram realizadas buscas por BLAST utilizando a consenso obtida para o DNAsat *pBuM* como *query.* Foram encontrados 545 *contigs* com *e-value* máximo de $10e^{-5}$, destes, apenas 341 foram analisados por possuírem um *hsp* maior que o tamanho da unidade de repetição. Os *contigs* em que foram encontradas cópias do *pBuM* variaram entre 200 e 45063 pb, embora apenas 17 *contigs* possuíam mais de 600 pb. Buscas por similaridade nos *scaffolds* montados resultaram em apenas seis *hits* com *e-value* significativo (inferior a $10e^{-6}$).

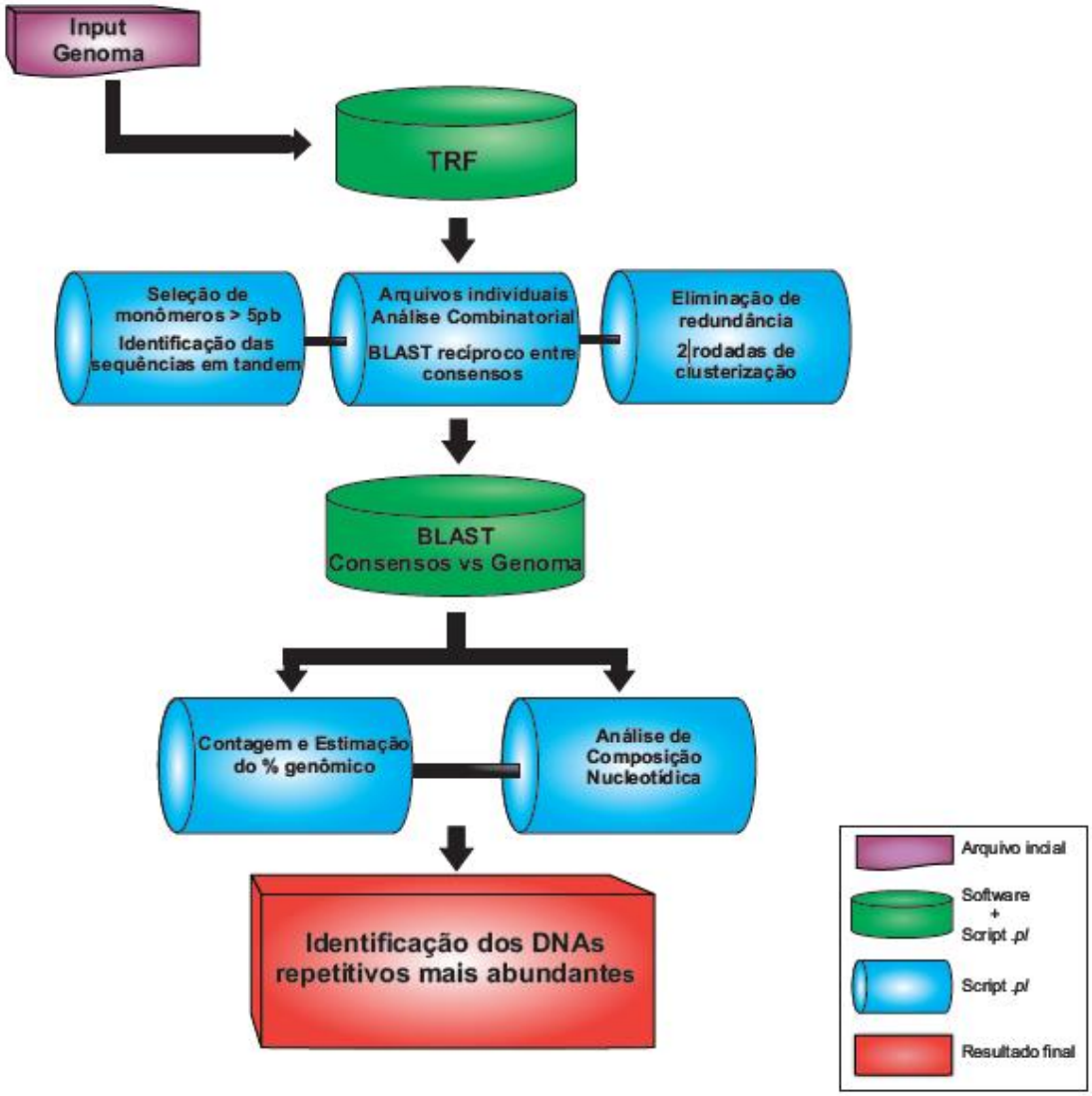

Figura 4. Workflow desenvolvido para a caracterização de sequências altamente repetitivas no genoma de D. buzzatii utilizando contigs montados gerados pela plataforma 454-Roche.

Foram identificadas um total de 838 *hsp* contendo sequências do *pBuM* presentes nos 341 *contigs*. No entanto, apenas 247 cópias apresentaram o tamanho de uma unidade de repetição completa de 189 pb. Também foram analisadas 132 cópias parciais compostas por mais de 120 pb.

A segunda família de DNA satélite abundante identificada em *D. buzzatii* apresentou um monômero de 198 pb, sendo denominada de *TR198*. A unidade de repetição do *TR198* não possui similaridade com nenhum DNA satélite ou elemento transponível descrito. Buscas por BLAST nos *contigs* do genoma montado de *D. buzzatii* resultaram em 171 *hits*

com *e-value* inferior a $10e^{-5}$, enquanto foram encontrados 51 *scaffolds* contendo sequências deste DNA satélite. Foram mineradas e caracterizadas mais de 200 cópias do *TR198,* sendo apenas 79 cópias completas.

As proporções genômicas de *pBuM* e *TR198* foram calculadas de acordo com o número de sequências obtidas, assumindo que o tamanho do genoma de *D. buzzatii* seja de 163,547,398 pb (Guillen e cols. 2015). Com essa metodologia, foi estimando que o DNA satélite *pBuM* representa 0,039 %, enquanto que o *TR198* representa 0,027 % do genoma (Tabela 2)

Tabela 2. Famílias de DNA satélite descritas no genoma de *D. buzzatii* utilizando o pipeline com 11 scripts de *BioPerl* desenvolvidos neste estudo.

Família de DNAsat	Monômero	% GC	Proporção Genômica (%)	Sequência Consenso	Distribuição Filogenética
pBuM	189	29	0.039	GCAAAAGACTCCGTCAATTAGAAAACA AAAAATGTTATAGTTTTGAGGATTAACCG GCAAAAACCGTATTATTTGTTATATGATTTC TGTATGGAATACCGTTTTAGAAGCGTCTTT TATCGTATTACTCAGATATATCTTAAGATTT AGCATAATCTAAGAACTTTTTGAAATATTC ACATTTGTCCA	*D. buzzatii cluster* species *D. mojavensis*
TR198	198	34	0.027	AAGGTAGAAAGGTAGTTGGTGAGATAA ACCAGAAAAAGAGCTAAAAACGGCTA AAAACGGCTAGAAAATAGCCAGAAAG GTAGATTGAACATTAATGGGCAAATGGA TGGATAAATAAGACTGGTCATCATCCAA TGAACAGAATCATGATTAAGAGATAGAA ATATGATTAGAAAGTAGGATAGAAAGGT TAGAAAG	*D. buzzatii*

.

Os resultados obtidos nesta etapa e organizados na Tabela 2 foram utilizados para compor os dados presentes no artigo de descrição do genoma de *D. buzzatii* (Guillen e cols. 2015) (Anexo 3).

5. Capítulo II – Artigo publicado na revista G3 Genes Genomes and Genetics: Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes

Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes

Leonardo G. de Lima[1], Marta Svartman[1], Gustavo C.S. Kuhn[1]

[1] Universidade Federal de Minas Gerais, Laboratório de Citogenômica Evolutiva, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Avenida Presidente Antônio Carlos, 6627 – Pampulha, 31270-901. Belo Horizonte, Brazil.

Abstract

Eukayote genomes are replete with repetitive DNAs. This class includes tandemly repeated satellite DNAs (satDNA) which are among the most abundant, fast evolving (yet poorly studied) genomic components. Here, we used high throughput sequencing data from three cactophilic *Drosophila* species, *D. buzzatii*, *D. seriema* and *D. mojavensis,* to access and study their whole satDNA landscape. In total, the *RepeatExplorer software* identified five satDNAs, three previously described (*pBuM*, *DBC-150* and *CDSTR198*) and two novel ones (*CDSTR138* and *CDSTR130*). Only *pBuM* is shared among all three species. The satDNA repeat length falls within only two classes, between 130-200bp or between 340-390bp. FISH on metaphase and polytene chromosomes revealed the presence of satDNA arrays in at least one of the following genomic compartments: centromeric, telomeric, subtelomeric or dispersed along euchromatin. The chromosomal distribution ranges from a single chromosome to almost all chromosomes of the complement. Fiber-FISH and sequence analysis of contigs revealed interspersion between *pBuM* and *CDSTR130* in the microchromosomes of *D. mojavensis*. Phylogenetic analyses showed that the pBuM satDNA underwent concerted evolution at both interspecific and intraspecific levels. Based on RNAseq data, we found transcription activity for *pBuM* (in *D. mojavensis*) and *CDSTR198* (in *D. buzzatii*) in all five analyzed developmental stages, most notably in pupae and adult males. Our data revealed that cactophilic *Drosophila* present the lowest amount of satDNAs (1.9% to 2.9%) within the *Drosophila* genus reported so far. We discuss how our findings on the satDNA location, abundance, organization and transcription activity may be related to functional aspects.

Key words:

Satellite DNA; cactophilic *Drosophila*, Centromeres; Telomeres; Concerted Evolution.

Introduction

The genomes of many organisms are replete with highly repetitive (>1,000 copies) tandemly repeated DNA sequences, commonly known as satellite DNAs (satDNAs) (Tautz 1993). Long and homogeneous arrays made of satDNA repeats are located in the heterochromatin (Charlesworth et al. 1994; Plohl 2012; Beridze 2013; Khost et al. 2016), but recent studies also revealed the presence of short arrays dispersed along the euchromatin (Kuhn et al. 2012, Brajkovic 2012, Larracuente 2014, Pavlek 2015). SatDNAs do not have the ability to transpose by themselves as transposable elements (TEs) do. However, there are some reported examples showing that TEs may act as a substrate for satDNA emergence and mobility (Dias et al. 2014; Mestrovic et al. 2015; Satovic et al. 2016).

The whole collection of satDNAs makes large portions (usually more than 30%) of animal and plant genomes (reviewed by Plohl et al. 2007). Although satDNAs do not code for proteins, they may play important cellular roles, including participation in chromatin packaging (Blattes et al. 2006; Fellicielo et al. 2015), centromere formation/maintenance (Rosic et al. 2014; Aldrup-MacDolnald et al. 2016) and gene regulation (Menon et al. 2014; Fellicielo et al. 2015; Urrego et al. 2017).

Despite their abundance, diversity and contribution to genomic architecture and function, our knowledge about several features of satDNAs is still limited. In the past decades, satDNAs have been mostly studied from a small sample of cloned repeats obtained by biased experimental approaches (usually by restriction digestion and/or PCR), isolated from one or few species. Experimental strategies for the identification of satDNAs were expensive, time-consuming and insufficient for the identification of the whole collection of satDNAs from any chosen genome.

Next-generation sequencing technologies have provided a revolution in the number of species with sequenced genomes, while new and efficient bioinformatic tools have been specifically developed towards genome-wide identification of repetitive DNAs. Consequently, we have now new tools and strategies to access the whole collection of satDNAs from a

given genome. For example, software tools known as *RepeatExplorer* have been successfully used for genome-wide characterization of repetitive DNAs from several animal and plant genomes, including those sequenced with less than 1x coverage (Barghini et al. 2014; Marques et al. 2015; Ruiz-Ruano et al. 2016; Zhang et al. 2017). This algorithm directly uses short next generation sequencing reads as rough material for the identification of repeats. Together with the results from similarity searches and abundance, the repeat families can be identified and classified.

Within the genus *Drosophila*, most studies on satDNA were conducted in *D. melanogaster* and in a few closely related species from the *melanogaster* group (e.g. Strachan et al. 1985; Kuhn et al. 2012; Larracuente et al. 2014; Garrigan et al. 2014; Jagannathan et al. 2016). The study of satDNAs of species distantly related to *D. melanogaster* are expected to broaden the understanding of this major fraction of the eukaryote genome. In this context, the *repleta* group is of particular interest. It contains at least 100 species that breed in cactuses in North and South America (Oliveira et al. 2012). Species from the *repleta* group are separated from the *melanogaster* group by more than 40My (Powell et al. 1997). Intense vertical studies in some species of this group revealed several aspects related to chromosome and genome evolution that have broad interest (e.g. Cáceres et al. 1999; Negre et al. 2005; Kuhn et al. 2009; Guillen et al. 2015).

At present, three *repleta* group species have available sequenced genomes: *D. mojavensis* (*Drosophila* 12 Genomes Consortium 2007), *D. buzzatii* (Guillen et al. 2015) and *D. seriema* (Dias et al. in prep). *D*. *buzzatii* and *D. seriema* belong to the *buzzatii* cluster, a monophyletic group of South American origin that contains seven species morphologically very similar and came from an radiation process dated at 6 Mya (Manfrin and Sene 2006; Oliveira et al 2012). *D. mojavensis* lives in the deserts and dry tropical forests of the southwestern United States and Mexico (Reed et al. 2007). The time since the split between

D. buzzatii and *D. mojavensis* has been estimated in 11 Mya (Oliveira et al. 2012; Guillén et al. 2015).

Previous studies in *D. buzzatii* and *D. seriema* conducted before the genomic era allowed the identification of three satDNA families. The first family, named *pBuM*, can be divided into two subfamilies according to its primary structure and size of the repeat units (Kuhn and Sene 2005). The *pBuM*-1 subfamily is comprised of *alpha* repeat units of approximately 190 bp, whereas the *pBuM*-2 subfamily consists of 370 bp composite repeat units called *alpha/beta,* each one consisting of an *alpha* (~190 bp) followed by a *beta* sequence (~180bp) of unknown origin. DNA hybridization data revealed pBuM-1 to be the major repeat variant present in *D. buzzatii* but pBuM-2 as the major repeat variant in *D. seriema*.

The second family, named *DBC-150*, consists of 150 bp long repeat units. This family is abundant in *D. seriema* but virtually absent in *D. buzzatii* (Kuhn et al. 2007). Finally, the third satDNA family, named *SSS13*9, with 139 bp long repeat units is abundant in *D. seriema* but absent in *D. buzzatii* (Franco et al. 2008). There is no significant sequence similarity among from *pBuM*, DBC-150 and *SSS139* satDNA families repeats, suggesting that they have independent evolutionary origins.

Three sequencing platforms (Sanger, 454 and Illumina) (Guillén et al. 2015) have been used to sequence the *D. buzzatii* genome, which became publicly available in 2015 (http://dbuz.uab.cat). In a preliminary approach, we used the Tandem Repeats Finder (TRF) software (version 4.04) (Benson 1999) to search for satDNAs with repeats longer than 50 bp in the *D. buzzatii* contigs. The two most abundant tandem repeat families identified were *pBuM*-1 (*alpha* repeats) and a novel family that we named *CDSTR198*, with 198 bp long repeat units (Guillén et al. 2015). However, in *D. melanogaster* and *D. virilis,* for example, several abundant satDNA families showed repeat units less than 10 bp long (Gall et al. 1971; Lohe et al. 1993). Therefore, a new satDNA screen is necessary in the *D. buzzatii* sequenced genome in order to look for the presence of small-size satDNA repeat motifs.

There are no detailed studies involving satDNAs in *D. mojavensis*. Melters *et al*. (2013) developed a bioinformatic pipeline to identify the most abundant tandem repeats from 282 selected sequenced genomes from animal and plant species, including some *Drosophila* species. A satDNA with 183 bp long repeat units was identified as the most abundant satDNA of *D. mojavensis*. Most recently, we showed that this satDNA actually belongs to the *pBuM*-1 satDNA subfamily (*alpha* repeats), previously described in *D. buzzatii* (Guillén et al. 2015).

Our group has recently sequenced the genome of *D. seriema* using the MiSeq platform (Dias et al. in preparation). The availability of three sequenced genomes (*D. buzzatii D. seriema* and *D. mojavensis*) provides an unprecedented opportunity to study the satDNA collection from each species and to compare them in a scale never possible before. We combined bioinformatic, phylogenetic and molecular cytogenetic tools to study the satDNA fraction from these three cactophilic *Drosophila* species. The resulting data are discussed in the context of satDNA genomic distribution, evolution and potential functional roles.

Material and Methods

Genomic data

The Illumina sequence reads from *D. buzzatii*, *D. mojavensis* and *D. seriema* used for identification of satDNAs were obtained from three different sources: *D. buzzatii* sequence reads (76x coverage) (available to download at http://dbuz.uab.cat) were generated by Prof. Alfredo Ruiz (Universitat Autònoma de Barcelona, Spain); *D. mojavensis* sequence reads (20x coverage) were generated by Prof. Bernardo de Carvalho (Universidade Federal do Rio de Janeiro, Brazil); and *D. seriema* (ERX2037878) sequence reads (20x coverage) were generated by our group (Dias et al. in prep).

Identification of satellite DNAs

Similarity-based clustering, repeat identification and classification were performed using *RepeatExplorer* (Novák et al. 2013) with whole-genome shotgun (WGS) *Illumina* reads from *D. buzzatii*, *D. mojavensis* and *D. seriema*. Initially, files containing all sequence reads from each species were uploaded (trimmed at 100 bp). The clustering analysis used *RepeatExplorer* default parameters. Clusters containing possible tandemly repeated satDNA families were identified based on the resultant graph-based clustering and then manually checked for the presence of tandem repeats using the Tandem Repeats Finder (TRF) software (version 4.04) (Benson 1999). Genomic proportion was calculated from the number of reads present in each cluster divided by the total number of reads. We searched for clusters with high graph density, which is a typical characteristic of satDNAs families (Novák et al. 2013). The *Dotlet* software (Junier and Pagni 2000) was also used to generate a scrutinized description of full length copies of each satDNA family.

Sequence and phylogenetic analysis

Multiple satDNAs sequences were aligned with the *Muscle* algorithm (Edgar 2004) of the MEGA5.05 software (Tamura et al. 2011), with manually optimization when necessary. MEGA5.05 was also used for the analysis of nucleotide composition and variability. Phylogenetic trees were constructed with the Neighbor Joining algorithm (Saitou and Nei 1987) of the MEGA program 5:05 (Tamura et al. 2011). The genetic distance between sequences was calculated using the "Tamura-Nei model" (Tamura and Nei 1993) after an analysis of best substitution model for the data on MEGA 5.05 (Tamura et al. 2011). Statistical evaluation of each branch of the tree was performed using analysis "bootstrap" (1,000 replicates).

Samples, DNA extractions, PCR amplifications, cloning and sequencing

For our experimental data we used DNA from the same sequenced strains: *D. buzzatii* (strain: ST01), *D. seriema* (strain: D73C3B) and *D. mojavensis* (strain: CI 12 IB -4 g8). DNA extraction of 30-50 adult flies was performed with the Wizard® Genomic DNA

Purification kit (Promega). PCR reactions consisted of an initial denaturation step of 94 ºC for 3 min, followed by 30 cycles of 94 ºC for 60 sec, 55 ºC for 60 sec and 72 ºC for 60 sec and then a final extension at 72 ºC for 10 min. The primers used for satDNA amplification are listed on Table S1. PCR products were excised from 1% agarose gels and purified with the Wizard SV Gel and PCR Clean-up System kit (Promega). After cloning with the pGEM-T-Easy cloning kit (Promega), recombinant plasmids were sequenced on the ABI3130 platform (Myleus Biotechnology).

***In situ* hybridization experiments**

Chromosome preparations, DNA fibers obtention, single and double-colour FISH and Fiber-FISH experiments were conducted as described in Kuhn et al. (2008). The probes labeled with digoxigenin-11-dUTP were detected with anti-digoxigenin FITC (Roche) and probes labeled with biotin-14-dATP were detected with NeutrAvidin-rhodamine (Roche). Chromosomes were stained with DAPI (4 ', 6-diamidino-2-phenylindole, dihydrochloride salt). The preparations were analyzed under an epifluorescence Zeiss Axiophot 2 microscope equipped with a CCD camera and the images were obtained with the AxioVision software (Zeiss). To determine the size of the DNA fibers, hybridization signals were measured according to the protocol described by Schwarzacher and Heslop-Harrison (2000).

Transcription Analysis

Total RNA-Seq data of *D. mojavensis* and *D. buzzatii* (st-1 strain) were those obtained by Guillen et al (2015). Briefly, RNA samples were extracted from 10-20 individuals from each of the four development stages (embryos, third-stage larvae, pupae, adult females and males), enriched for mRNA by poly-A tail selection and sequenced by Illumina, generating ~100 bp reads (see Guillen et al. 2015 for details). All reads were aligned against consensus sequences representing the *pBuM* and *CDSTR198* families from *D. buzzatii* and *pBuM* and *CDSTR*130 from *D. mojavensis* with the Bowtie2 software (Langmead and

Salzberg 2012) incorporated into the usegalaxy.org server (Afgan et al. 2016). The mapped reads were normalized by the RPKM method (reads per kilobase per million mapped reads; Mortavazi et al. 2008).

Results and Discussion

Cactophilic *Drosophila* Repetitive DNAs: general aspects

The *RepeatExplorer* graphic representation containing all identified repetitive DNA clusters in *D. buzzatii, D. seriema* and *D. mojavensis* and their genome proportion (%) is shown in Figures S1-S3. Most clusters making more than 0.01% of the genome could be classified into established groups of repetitive elements, such as TEs, satDNAs or rDNA sequences (Figure 1; Tables S2-S4).

The satDNA genomic contribution is similar in the three species: ~1.9% in *D. buzzatii,* ~2.9% in *D. seriema* and ~2.5% in *D. mojavensis*. The genomic contribution of the classified TEs is on average 5.4 x higher: ~12% in *D. buzzatii*, ~18% in *D. seriema*, and ~11% in *D. mojavensis.* Rius et al. (2016) have recently estimated the TE content of *D. buzzatii* and *D. mojavensis* using the same genomic sequences used in this work, but with a different methodology) and found that TEs represent ~11% of the *D. buzzatii* and ~15% of the *D. mojavensis* genomes.

The genomic contribution of the different TE orders (TIR-transposons, Helitrons, LTR-retrotransposons and Non-LTR retrotransposons) differs among the three species (Figure 1). TIR-transposons are the most abundant TEs in the *D. buzzatii* genome (3.85%); in *D. seriema* LTR-retrotransposons (6.8%) are the most abundant and in *D. mojavensis*, Helitrons are the most abundant TE elements (3.25%). Conversely, Rius et al. (2016) described Helitrons as the most abundant TEs in the *D. buzzatii* and *D. mojavensis* genomes. Interestingly, the genomic contribution of LTR-retrotransposons in *D. seriema* (6.8%) is at least two times higher than in *D. buzzatii* (2.9%) or in *D. mojavensis* (2.4%). The contribution of unclassified repetitive elements is also considerably higher in *D. seriema*

(18%) than in the other two species (11% and 12%). These results suggest a recent burst of repetitive elements in *D. seriema*.

Satellite DNA landscape in the three cactophilic *Drosophila* species

We identified only two previoulsy described satDNA families in *D. buzzatii,*. The *pBuM*-1 satDNA (Kuhn and Sene 2005) with 189 bp long *alpha* repeats is the most abundant, representing 1.7%. The second is *CDSTR198* (Guillen et al. 2015), with 198 bp long repeats and representing 0.2% of the genome. These genomic contributions revealed by *RepeatExplorer* are higher than those obtained by our first contig-based approach, most notably for *pBuM*-1 (0.04% for *pBuM*-1 and 0.03% for *CDSTR198*; Guillen et al. 2015). The organization of satDNAs, made of several tandem repeats with high DNA sequence similarity, imposes a huge limitation for assembly computer programs. Consequently, it is very likely that the bulk of *pBuM* and *CDSTR198* satDNA repeats of *D. buzzatii* were omitted from the contigs used in our previous approach. Accordingly, although still low (see discussion below), we consider the values obtained in the present work as the most reliable ones.

We detected four satDNAs in *D. seriema*. The *pBuM*-2 satDNA with ~340-390 bp long *alpha/beta* repeat units (Kuhn and Sene 2004) is the most abundant, representing 1.93% of the genome. The second satDNA is DBC-150 (Kuhn et al. 2007), with ~110-150 bp long repeat units and representing 0.8% of the genome. The third satDNA is a novel one and was named *CDSTR*138, with 138 bp long repeat units and representing 0.23% of the genome. The fourth satDNA is *CDSTR198*, which isshared with *D. buzzatii*, but represents only 0.02% of the *D. seriema* genome.

The SSS139 satDNA, with 139 bp long repetition units was previously described in *D. seriema* (Franco et al. 2008). In the *RepeatExplorer* output, we found sequences homologous to SSS139 in the 10th most abundant repeat cluster, representing 0.5% of the

genome. However, detailed sequence analysis revealed that this cluster is not made of tandem repeats. Instead, most sequences correspond to a ~30 bp SSS139 inverted fragment interrupted by a region variable both in size and identity, followed by a ~ 120 bp SSS139 sequence in direct orientation. Interestingly, these variable regions or the SSS139 sequences themselves showed no similarity to any TE or satDNA family previously described. Therefore, further studies will be necessarily for elucidating the nature of the SSS139 repetitive elements.

We found two satDNAs in *D. mojavensis*. The most abundant is a novel one, which we named *CDSTR*130, with 130 bp long repeat units and representing 1.63% of the genome. It is worth noting, however, that RepBase identified these sequences as a Long Terminal Repeats (LTR) BEL3_DM-I element described in *D. mojavensis* (Jurka 2012). This LTR has been characterized from *D. mojavensis* scaffold 5562 (nucleotide positions 8682 to 13043 bp). However, the scrutinized analysis of 100 BEL3-DM insertions on the *D. mojavensis* genome showed that the 130 bp tandem repeats are not part of the LTR, but only flank the element in the scaffold 5562 (Figure 2). The identification of *CDSTR*130 as a satDNA highlights the importance of manual curation of the automated output provided by *RepeatExplorer*. It also explains why Melters et al. (2013) did not identify *CDSTR*130 as the most abundant tandem repeat family in the *D. mojavensis* genome.

The second most abundant satDNA identified in *D. mojavensis* is the *pBuM*-1 variant from the *pBuM* family (shared with *D. buzzatii and D. seriema*), with 185 bp long repeats and representing 0.86% of the genome. This satDNA has been previously identified as the most abundant tandem repeat family of *D. mojavensis* by Melters et al (2013).

The main features of the satDNAs identified above are summarized in Table 1 and a list containing consensus sequences from all the new satellites described in the present work can be seen in Figure S4.

Cactophilic *Drosophila* species present the lowest satDNA content within the genus

In most analyzed *Drosophila* species, the satDNA proportion fall within the range of between 15-40% (Bosco et al 2007; Craddock et al. 2016). We found that the *pBuM* and *CDSTR*130 satDNAs represent only 2.5% of the *D. mojavensis* genome. Our result, obtained from the analyses of sequence reads using *RepeatExplorer*, was very close to the 2% satDNA contribution estimated by Bosco et al. (2007) using flow cytometry. In addition, we also found low amounts of satDNAs in the genomes of the other two cactophilic *Drosophila*: 1.9% for *D. buzzatii* and 2.9% for *D. seriema*. The additional 1% of the *D. seriema* in relation to *D. buzzatii*is probably represented by sequences located in the microchromosome of *D. seriema*, which is larger than that of *D. buzzatii* and also contains a higher amount of satellites (pBuM-2 and DBC-150) when compared to the other chromosomes (Figure 9; Kuhn et al. 2007, 2009). Our data revealed that cactophilic *Drosophila* present the lowest amount of satDNAs within the *Drosophila* genus reported so far. On the other hand, the estimated contribution of repetitive DNAs (satDNA+TE+unclassified repeats) in the three cactophilic *Drosophila* (14%-27%) is not atypical for the genus (*Drosophila* 12 Genomes Consortium 2007; Craddock et al. 2016). Future studies focusing on satDNAs of more populations and species of the *repleta* group are expected to shed light on whether the low satDNA content in cactophilic *Drosophila* is a result of selective constraints or historical events.

Preferential satDNA repeat lengths in cactophilic *Drosophila*

SatDNA repeats in the three studied cactophilic *Drosophila* have lengths of 130-200 bp or between 340-390 bp. In order to confirm this result, we ran *RepeatExplorer* with sequence reads from *D. melanogaster* where satDNA repeats less than 10bp are abundant. *RepeatExplorer* correctly identified them as the most abundant repetitive DNAs of *D. melanogaster* (Table S5). Therefore, we concluded that the preferential lengths for satDNA repeats in the three cactophilic *Drosophila* are not an artifact generated by *RepeatExplorer*.

Interestingly, satDNA repeats described before the genomic era in many plant and animal species (including *Arabidopsis*, maize, humans and many insect species) typically show basic repeat units 150-180 or 300-360bp long (Henikoff et al. 2001; Heslop-Harrison et al. 2003). Similar repeat-length patterns have been confirmed with recent genome-wide analysis of tandem repeats in other organisms. For example, Pavlek et al. (2015) showed that the most abundant tandem repeat families in the beetle *Tribolium castaneum* present repeat lengths either around ~170 bp or around ~340 bp long. It is difficult to explain such preferential repeat lengths by chance. On the other hand, it is striking that these two peak units closely correspond to the length of DNA wrapped around one or two nucleosomes.

It has been hypothesized that satDNA length could play a critical role in DNA packaging by favoring nucleosome positioning (or phasing) that in turn leads to condensation of certain genomic regions, such as the heterochromatin (Fitzgerald et al. 1994; Henikoff et al. 2001). Accordingly, the preferential lengths observed in the satDNA from cactophilic *Drosophila* could be selectively constrained by a possible role in chromatin packaging.

Satellite DNA candidates for centromeric function

The centromeres of most plant and animal species are composed of long arrays of tandemly repeated satellite DNAs (Plohl et al. 2014). There is increasing evidence to a role for satDNA in centromeric function by providing motifs for centromeric-protein binding, e.g. CENP-B box in alphoid human satDNA (Ohzeki et al. 2002), and/or by producing RNA transcripts that are necessary to centromere/kinetochore assembly (Gent and Dawe 2012; Rosic et al. 2014). On the other hand, centromeric satDNAs may differ greatly even between closely related species. In fact, there are several examples supporting the observation that satDNA is one of the most rapidly evolving components of the genomes. Therefore, the identification of the most likely candidate for centromere function in a species is a task that in most cases has to be performed on a case-by-case basis.

Based on data collected from several animal and plant genomes, Melters et al. (2013) suggested that the most abundant tandem repeat of a genome would also be the most likely candidate for centromeric location and function. In order to test this hypothesis, we investigated by FISH the chromosomal location of all satDNAs identified in the three cactophilic *Drosophila* sampled in the present study.

All three species share the same basic karyotype (2n=12) consisting of four pairs of telocentric autosomes, one pair of microchromosomes and one pair of sex chromosomes (Baimai et al. 1983; Kuhn et al. 1996;Ruiz et al. 1990). Heterochromatin is located in the centromeric region of all four telocentric chromosomes, along the whole microchromosomes and Y chromosome and covering approximately 1/3 of the proximal region of the X chromosome.

We identified the *pBuM*-1 alpha repeats as the most abundant satDNA of *D. buzzatii*. In a previous study, Kuhn et al. (2008) showed by FISH on mitotic chromosomes that *pBuM*-1 alpha repeats are located in the centromeric heterochromatin of all chromosomes except the X. In order to further investigate the chromosomal location of *pBuM*, we also hybridized a *pBuM-1* probe to the polytene chromosomes. In these chromosomes, the centromeric heterochromatin is underreplicated and forms a dense central mass in the chromocenter - a region where the centromeres of all chromosomes bundle together. We observed that the *pBuM-1* repeats are restricted to the chromocenter region (Figure 3a), therefore confirming their centromeric location. The second most abundant satDNA in *D. buzzatii* is *CDSTR198*, which was mapped by FISH in terminal and interstitial locations on metaphase chromosomes (these results are detailed below). Therefore, the most abundant satDNA of *D. buzzatii*, i.e., *pBuM*, is the one showing centromeric location in most chromosomes.

In *D. seriema*, the most abundant satDNA identified was *pBuM*-2 and the second most abundant was *DBC-150*. Previous studies showed that *pBuM*-2 is located on the

centromeric regions of chromosomes 2, 3, 4 and 5 and on the telomeric regions of chromosome 6 (Kuhn et al. 2008). *DBC-150* was found exclusively on the centromeric region of chromosome 6 (Kuhn et al. 2007). *CDSTR138,* the new satDNA described herein*,* is the third most abundant tandem repeat of this species and was mapped by FISH at the centromeric region of chromosomes 2, 3, 4 and 5 in mitotic chromosomes (Figure 4b). The centromeric location was also confirmed after FISH on polytene chromosomes, where no hybridization signals were observed outside the chromocenter (Figure 3a). The fourth identified satDNA in *D. seriema*, *CDSTR198,*showed no hybridization signal after FISH on mitotic chromosomes, confirming that it has very low copy number in this species (in contrast to *D. buzzatii*). However, we detected a few *CDSTR198* repeats in the euchomatin after FISH on polytene chromosomes (Figure 3b; see below)*.* Therefore, all three most abundant satDNAs of *D. seriema* are part of the centromeric region of most chromosomes.

CDSTR130 was identified as the most abundant satDNA in *D. mojavensis*, FISH on mitotic chromosomes showed that *CDSTR130* repeats are located at the centromeric region of all autosomes and the X chromosome (Figure 4d). The second most abundant satDNA is *pBuM*-1, which covered the microchromosome (chromosome 6) almost entirely (Figure 4d). Therefore, both *pBuM*-1 and *CDSTR*130 are abundant in chromosome 6. However, given the size and dot-like morphology of this chromosome in this species, it is not possible to determine which one shows centromeric location. The analysis of the polytene chromosomes showed that the two satDNAs co-localize in the chromocenter region (Figure S5).

Based on the collection and chromosome distribution of the satDNAs discussed herein, the centromeric regions of the X chromosome of *D. buzzatii*, of the X and Y of *D. seriema* or of the Y of *D. mojavensis* are not composed of satDNAs. Some centromeres described in plants and animals are composed of transposable elements (reviewed by Plohl et al. 2014). In *Drosophila*, DINE-1 elements (helitrons) are one of the most abundant types of transposable elements (Yang and Barbash 2008). Kuhn and Heslop-Harrison (2011) and Dias et al. (2015) showed by FISH on mitotic chromosomes that these elements are highly

enriched in the sex chromosomes (including the centromeric regions) in the three analyzed species from the *repleta* and *virilis* groups. It is possible that these DINE-1 elements are the main components of the centromeres of the sex chromosomes of cactophilic *Drosophila* species.

According to *RepeatExplorer*, the genomic proportion of satDNA in *D. mojavensis* (*CDSTR130*+*pBuM*) is 2.5% (Table 1). This value is very close to the 2% satDNA contribution estimated by Bosco et al. (2007) using flow cytometry in the same species. According to the authors, if we split the ~2% satDNA evenly among the *D. mojavensis* chromosomes that would result in ~430 kb for each centromere. As noted by the authors, this value is also very close to what is considered as the minimum amount of centromeric DNA (420kb) needed to fulfill centromeric function in *Drosophila* (Sun et al. 1993). In this context, Bosco et al (2007) emphasized that it would be valuable to identify the centromeric satDNA of *D. mojavensis* and other *Drosophila* species to investigate whether they agree with the ~420kb limit observed in *D. melanogaster*.

In the present work, we found that *pBuM* and *CDSTR*130 are the main centromeric components of *D. buzzatii* and *D. mojavensis*. According to previous estimates, the male genome size of *D. buzzatii* and *D. mojavensis* is around 170 Mb (Gregory and Johnston 2008; Romero-Soriano et al. 2016). Accordingly, we calculated that the bulk of centromeric satDNA in *D. buzzatii* is 2.9 Mb and in *D. mojavensis*, 2.8 Mb. If we split these values equally between the number of centromeres (= 6), each centromere will have ~480 kb of centromeric DNA in *D. buzzatii* and ~460 kb in *D.mojavensis*. These suggests cactophilic *Drosophila* have centromeric sizes roughly470 kb on average, a value close to the suggested limit of 420 kb necessary for a functional centromere in *Drosophila* (Sun et al. 1993).

New insigths on *pBuM* distribution and evolution

According to previous data on the distribution of *pBuM*-1 *alpha* and *pBuM*-2 *alpha/beta* repeats in the phylogeny of *Drosophila* species from the *buzzatii* cluster (*repleta* group), it was proposed that the ancestral state of the *pBuM* satDNA family consisted of *alpha* tandem repetition units around 190bp long. The *alpha/beta* repeats would have been originated subsequently from an insertion of a non-homologous sequence of 180 bp (*beta*) in an *alpha* array, resulting in a composite *alpha/beta* repeat unit that also became abundant and tandemly organized (Kuhn and Sene 2005).

We found only *alpha* repeats in the genome of *D. mojavensis*, which is consistent with the hypothesis that *alpha* repeats represent the ancestral state of the *pBuM* family. According to current estimates, the split between the *buzzatii* and *mojavensis* clusters occurred around 11 Mya (Oliveira et al. 2012; Guillén et al. 2015), which would be the minimum age for the origin of the *pBuM* family.

In *D. seriema*, we detected only *pBuM-2* repeats, which agrees with previous DNA hybridization data (Kuhn and Sene 2005) suggesting that *pBuM-2* is the only *pBuM* subfamily present in this species. The split between *D. buzzatii* and *D. seriema* was estimated to have happened around 3Mya (Franco et al. 2010). Therefore, in the last 3My, it seems that there was a complete turnover from *pBuM-1* to *pBuM-2* repeats in the genome of *D. seriema*.

According to our FISH experiments on mitotic and polytene chromosomes, *pBuM* repeats are restricted to the heterochromatic regions. However, BLAST on the assembled genome (Freeze 1 Scaffolds) of *D. buzzatii* revealed fragments of *pBuM-1* repeats on three scaffolds (1, 88 and 90) that were mapped to the euchromatin from chromosomes 2, 5 and X (see Guillén et al. 2015 for exact location of scaffolds). The three observed *pBuM*-1 euchromatic loci contain either a partial *pBuM-1* repeat (less than 189 bp) or at most two partial *pBuM-1* tandem repeats (less then < 300bp), and such small sizes were probably the reason they were undetected in our FISH experiments. The analysis of flanking sequences did not show evidence that these euchromatic *pBuM-1* sequences could be integral parts of

transposable elements and the mechanism(s) responsible for their presence on euchromatin are currently unknown.

Previous phylogenetic analyses of *pBuM* repeats in *D. buzzatii* and *D. seriema* showed that these repeats have been evolving according to the concerted evolution model (Kuhn and Sene 2005). In other words, repeats within each species are more similar to each other than to repeats between species. In order to test whether *pBuM* also evolved in concert in *D. mojavensis*, we constructed a NJ tree with all *pBuM* repeats extracted from *D. buzzatii*, *D. seriema* and *D. mojavensis* (Figure 5). The NJ tree revealed *pBuM* repeats from each species allocated in species-specific branches, indicating that *pBuM* has been evolving in a concerted manner in the last 11Mya.

The presence of *pBuM* in the non-recombining Y allowed independent homogenization

In a previous report, the analysis of 63 *pBuM*-1 *alpha* repeats from *D. buzzatii* revealed very low levels of inter-repeat variability (4.2% on average), indicating that, despite multiple chromosomal location, *pBuM* arrays have been efficiently homogenized at the intraspecific level (Kuhn et al. 2003). However, one repeat (Juan/4) showed atypical levels of nucleotide divergence in comparison to the remaining repeats (22% on average). Kuhn et al. (2003) suggested that this repeat may belong to another, less abundant, *pBuM* subfamily.

In the present work, we retrieved a sample of 247 *pBuM-1* repeats from the sequenced genome of *D. buzzatii* and used them to construct a NJ tree. The resulting tree split the repeats into two main branches (Figure 6). The major one, containing 194 repeats, contains the “typical” *pBuM-1* repeats, described in Kuhn et al. (2003). The second minor branch, with 53 repeats, contains “Juan/4-like” *pBuM-1* repeats. Between the two groups, the nucleotide difference is 24.2%.

These data are consistent with the hypothesis of two *pBuM* subfamilies being present in the *D. buzzatii* genome. Herein, we will name them as *pBuM*-1a (typical) and

pBuM-1b ("Juan/4-like"). All the data generated so far about *pBuM* from *D. buzzatii* (including chromosomal location) concern the typical *pBuM*-1a repeat variant. There are several diagnostic nucleotide substitutions that allow discrimination between *pBuM* repeats from these two subfamilies. Such a situation allowed us to design oligonucleotides to specifically amplify *pBuM*-1b repeats by PCR for probe preparation. We then performed double-FISH with *pBuM*-1a and pBuM-1b on *D. buzzatii* mitotic chromosomes. The *pBuM*-1a probe showed the same multichromosomal distribution as described before. However, the *pBuM*-1b probe hybridized specifically to the Y chromosome (Figure 4a).

According to the model of concerted evolution, intraspecific homogenization of repeats occurs by recombination events such as unequal crossing over and gene conversion (Dover1982; Dover and Tautz 1986). There is also some evidence suggesting that different arrays on the same or in different chromosomes may experience independent homogenization for arrays- or chromosomal-specific repeat variants (i.e. intragenomic concerted evolution) (Kuhn et al. 2012; Larracuente2014; Khost et al. 2016). In this context, it is expected that arrays with tandem repeats on non-recombining chromosomes, such as the Y, would be specially subjected to independent homogenization. This is most likely the reason for the existence of a different *pBuM* subfamily (*pBuM*-1b) on the Y chromosome of *D. buzzatii*. Furthermore, empirical and experimental data showed that low recombination is expected to increase inter-repeat variability (Stephan and Cho 1994; Navajas-Pérez et al. 2006; Kuhn et al. 2007). In fact, *pBuM*-1a repeats had a nucleotide difference of 12%, while the *pBuM*-1b repeats (restricted to the Y chromosome) showed a higher variability of 17%.

The *CDSTR198* satDNA shows terminal and dispersed distribution

The *CDSTR198* satDNA was found in *D. buzzatii* and *D. seriema*, but with marked quantitative differences (0.23% in *D. buzzatii* and 0.02% in *D. seriema*). FISH on *D. buzzatii* mitotic chromosomes revealed that this satDNA is located in the terminal regions of chromosomes 2, 3, 4, 5 and X but also spread along euchromatic regions (Figure 4a). FISH

on polytene chromosomes of the same species revealed strong hybridization signals in the telomeric regions of chromosomes 2, 5 and X, and in subtelomeric regions of chromosomes 3 and 4 (Figure 3a). Moreover, we detected the presence of *CDSTR198* repeats along euchromatic regions of all chromosomes, except on the microchromosome. We found the highest number of *CDSTR198* euchromatic signals concentrated in chromosomes 2 and 5 (Figure 3a). Similar results were also obtained by an overall analysis of 37 *CDSTR198* euchromatic arrays present in the *D. buzzatti* assembled genome (Table S6). Interestingly, this analysis showed an equal number of euchromatic arrays present on chromosomes 2 and 3 (11 arrays each), followed by chromosomes 4 and 5 (six arrays each). The fewer euchromatic arrays found in the *D. buzzatii* genome may result from the computational challenge of repetitive element assembly (Treangen and Salzberg 2012), reinforcing the need of hybridization experiments of satDNA families spread throughout euchromatin. In line with this, it is relevant to suggest that some *CDSTR198* arrays identified by FISH may be absent on assembled genomes. FISH on polytene chromosomes of *D. seriema* showed *CDSTR198* located only in a few euchomatic sites (Figure 3b).

In contrast to transposable elements, satDNAs do not have the ability to transpose by themselves. However, there are some reported examples showing that TEs may act as a substrate for satDNA emergence and mobility (Dias et al. 2015; Mestrovic et al. 2015; Satovic et al. 2016). We created a database containing the 500bp sequences immediately before and after each *CDSTR198* array (37 in total; Table S6) found in the assembled scaffolds of *D. buzzatii.* Comparative analysis of all flanking sequences did not show association to a specific TE or TE family or to any other specific sequence common to all arrays. These results raise the question about the dispersion mechanism of *CDSTR198* in the *D. buzzati* genome.

Tandemly repeated sequences may undergo small recombination events involving copies of the same array in the same orientation. These events may result in the formation of

extrachromosomal circular DNAs (*eccDNAs*) (Cohen and Segal 2009). The occasional presence of a replication initiating region may provide further amplification and new *eccDNA* copies. Apparently, these *eccDNAs* can be inserted again into the genome by recombination. This mechanism was proposed to explain the dispersion of copies of the *satDNA* TCAST2 in *Tribolium castaneum* (Brajkovic et al. 2012), as well as of the *D. melanogaster 1.688* satDNA (Cohen and Segal 2009), which also show an euchromatic dispersed distribution (Kuhn et al. 2012). In order to test this hypothesis it would be interesting to look for the presence of *eccDNA* containing *CDSTR198* repeats in *D. buzzatii*.

CDSTR198* satDNA may contribute to telomeric function in *D. buzzatii

Unlike most eukaryotes, *Drosophila* telomeric regions are maintained by a sequence complex organized in three subdomains: (i) arrays of TEs (Het-A/TART) responsible for maintaining telomeric sequences; (ii) telomere-associated sequences (TAS), formed by complex repetitive sequences, usually satDNAs, and (iii) a protein complex HOAP required for telomere stability (Silva-Sousa et al. 2012). Although the structure of telomeres is conserved among all *Drosophila* species, the TEs and TAS sequences are highly variable even among phylogenetically close species (Villasante et al. 2007). Based on the widespread presence of TAS in *Drosophila* and other species (including humans), Biesmann et al. (2000) proposed that homologous recombination between terminal satDNA repeats could have been an “ancient” mechanism for telomere extension. Today, TAS regions probably function as a buffer zone between the telomeres and internal chromosome domains (Sharma and Raina 2005).

We could not identify conserved domains for telomeric Het-A and TART TEs in the sequenced genome of *D. buzzatii*, even though these TEs were described in *D. mojavensis* and *D. virilis* (Villasante et al. 2007). Similarly, a recent screening of the *D. buzzatii* sequenced genome for the whole TE content did not identify Het-A or TART elements (Rius et al. 2016). The apparent absence of Het-A and TART in *D. buzzatii* may be related to the

high evolutionary rate of these sequences (Villasante et al. 2007). Alternatively, there may be a different mechanism for telomere elongation operating in this species.

The *CDSTR198* satDNA is located in the telomeric and subtelomeric regions of five (out of six) chromosomes of *D. buzzatii* (Figures 3a; 4b). The presence of *CDSTR198* in the telomeres associated with the apparent absence of Het-A and TART sequences open the possibility that *CDSTR198* plays a role in telomere elongation through a recombination-based mechanism (e.g. unequal crossing over). Although not described in *Drosophila*, tandem repeat sequences are responsible for maintaining telomeres in the dipterous genus *Chironomus* (Lopez et al. 1996).

It is important to mention that a similar scenario described herein for the *CDSTR198* of *D. buzzatii* was previously reported for *D. virilis*, which belongs to the *virilis* group. In this non-cactophilic species, the terminal location of the *pvB370 satDNA* associated with the absence of telomere transposons led Biesmann et al. (2000) to propose the involvement of this satDNA in telomere elongation. However, TART-like and HeT-like elements were later described in the terminal regions of *D. virilis,* opening the possibility that these elements also participate in telomeric elongation in this species (Casacuberta et al. 2003; Pardue et al. 2005).

pBuM and _CDSTR_130 show regions of interspersed distribution in the microchromosomes

FISH with *CDSTR*130 and *pBuM* probes on *D. mojavensis* mitotic chromosomes revealed that these two satDNA colocalize on the microchromosome. In order to further investigate how these two satDNAs are organized we performed double FISH experiments on extended DNA fibers. We observed strong hybridization signals in fibers showing *CDSTR130* long arrays followed by *pBuM* long arrays (Figure 7a). However, in some DNA

fibers hybridization signals indicated an interspersed organization of both satDNAs (Figure 7b). These results were also confirmed in the analysis of *D. mojavensis* assembled contigs (Figure 7c). For example, the contig 2999 (AAPU01002998.1) is composed of 4,435 bp of *CDSTR130* copies adjacent to a *pBuM* array of 7,716 bp. In the contig 4,375 (AAPU01004374.1) we observed different arrays of *pBuM* and *CDSTR130* interspersed with each other (Figure 7c).

Non-homologous satDNAs located in the same chromosome region are usually organized in separate arrays (e.g. Shiels et al. 1997; Lohe et al. 1993; Sun et al. 2003). However, there are some reports showing interspersion of repeats from different satellites (e.g. Zinic et al. 2000; Alkhimova et al. 2004; Wei et al. 2014). It has been suggested that interspersion between repeats may give rise to new higher order repeat structures (Mravinac and Plohl 2007; Wei et al. 2014). In a previous study conducted in cactophilic *Drosophila* species, Kuhn et al. (2009) showed high levels of interspersion between *pBuM* and DBC-150 in at least two species of the *buzzatii* cluster (*D. gouveai* and *D. antonietae*). Interestingly, such pattern was also observed in the microchromosomes. According to Kuhn et al (2009), interspersion of repeats from non-homologous satellites in the microchromosomes could be related to the peculiar characteristics of these chromosomes, such as highly heterochromatic nature and low content of genes, which could allow a more flexible interplay between repetitive elements without deleterious effects.

Differential transcription of cactophilic *Drosophila* satDNAs

SatDNAs do not code for proteins and have been traditionally viewed as “junk DNAs”. However, there is a growing number of studies showing satDNA transcription activity from yeast to mammals and the biological function of these transcripts has now started to be appreciated. For example, satDNA transcripts were shown to be involved in heterochromatin assembly, kinetochore formation and gene regulation (reviewed by Biscotii et al. 2015; Ferreira et al. 2015). Moreover, transcription of satDNAs is usually gender or stage specific

and is often associated with differentiation and development (Usakin et al. 2007; Pecinka et al. 2010).

Herein, we investigated whether the satDNAs that we analyzed are transcribed by mapping the satDNA consensus sequences on the available RNA-seq data from *D. buzzatii* and *D. mojavensis* (Guillen et al. 2015; Rius et al. 2016). Read counts were calculated for embryos, third-staged larvae, pupae and for male and female adult carcasses (Figure 8) (See methods).

Our analysis did not identify transcripts from the most abundant satDNAs in the genome of *D. buzzatii* and *D. mojavensis*, *pBuM* and *CDSTR*130, respectively. As discussed previously, both are the main candidates for centromeric function in these species. This result was unexpected because previous studies in *Drosophila melanogaster* showed that centromeric satellite RNAs in the form of long polyadenylated products play an important role in the formation of the kinetochore (Topp et al 2004; Chan et al. 2012; Rosic et al. 2014). However, our results do not exclude the possibility that *pBuM* and *CDSTR*130 are transcribed. In this case, the absence of satDNA transcripts may be related to the methodology used for RNA extraction that preferentially captures poly(A) sequences. For example, satDNA transcripts of *D. melanogaster* involve ncRNAs that do not have poly(A) tails (Usakin et al. 2007).

Conversely, in all five analyzed tissues we detected transcripts derived from the *CDSTR198* satDNA of *D. buzzatii* and from the *pBuM* satDNA of *D. mojavensis.* In both cases, the transcripts were particularly abundant in tissues from pupae and males. Interestingly, these two satDNAs are located in different genomic environments: while *CDSTR198* arrays are located at several euchromatic loci (including some close to genes; Table S7) in several *D. buzzatii* chromosomes, pBuM is exclusively located in the heterochromatic microchromosome of *D. mojavensis*. Future studies will be needed to address whether these transcripts participate in chromatin modulation and/or if they affect the

transcription of neighboring genes, as observed for satDNA transcripts of *Drosophila* and other organisms (Menon et al. 2014; Fellicielo et al. 2015).

Figures Legends

Fig. 1: Estimated repetitive DNA abundance in three cactophilic *Drosophila* species.

Fig. 2: Schematic representation of the BEL3-DM-I transposable element present on RepBase, which is flanked by CDSTR130 satDNA arrays. Blue arrows represent the undescribed 185 bp long terminal repeat of the BEL3-DM element.

Fig. 3: FISH on polytene chromosomes of *D. buzzatii* (A) and (B) *D. seriema* using satDNA probes for *pBuM* (red) and *CDSTR198* (green) (Arrowheads indicate telomeric regions).

Fig. 4: FISH on mitotic chromosomes using satellite DNA probes. (A) *pBuM-1a* (red) and pBuM-1b (green) satDNA probes on *D. buzzatii;* **B.** *pBuM-1a* (red) and *CDSTR198* (green) probes on *D. buzzatii;* **C.** *CDSTR138* (red) on *D. seriema* (**D)** *CDSTR130* (green) and *pBuM* (red) probes on *D. mojavensis.*

Fig. 5: NJ tree containing a sample of *pBuM* repeats extracted from the sequenced genomes of *Drosophila buzzatii* (green), *D. seriema* (blue) and *D. mojavensis* (red). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

Fig. 6: NJ tree of *pBuM* satDNA repeats retrieved from the *D. buzzatii* assembled genome and previously described on Kuhn et al. (2003) Colored braches evidence Y chromosome specific arrays (yellow) when compared to autosomal arrays (green). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

Fig. 7: A-B FISH with *CDSTR*130 (green) and pBuM (red) probes onto extended DNA fibers of *D. mojavensis*. (C) Schematic representation of *CDSTR130* and *pBuM* organization found on *contigs Ctg01_2999(*AAPU01002998.1) and *Ctg01_4375* (AAPU01004374.1 retrieved from the *D. mojavensis* assembled genome.

Fig. 8: Transcription profile of satDNA families in *D. buzzatii* (A) and *D. mojavensis* (**B)** on five different developmental stages. Counts were normalized to one million reads.

Fig. 9: Representative ideogram showing the chromosomal localization of all satDNAs identified in *D. buzzatii*, *D. seriema* and *D. mojavensis*.

Table Legends:

Table 1. Main features of satellite DNA families present on *D. buzzatii*, *D. seriema* and *D. mojavensis* genomes.

Supplementary Material

Supplementary Figures Legends:

Fig. S1. Repetitive clusters (n=122) in *D. buzzatii* identified by RepeatExplorer after clusterization of 270366 reads. Together, these clusters represent 14.7% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-1 and CDSTR198 satellite DNAs are indicated.

Fig. S2. Repetitive clusters (n=328) in *D. seriema* identified by RepeatExplorer after clusterization of 526010 reads. Together, these clusters represent 26.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-2, DBC-150 and *CDSTR138* satellite DNAs are indicated.

Fig. S3. Repetitive clusters (n=217) in *D. mojavensis* identified by RepeatExplorer after clusterization of 323342 reads. Together, these clusters represent 14.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The *CDSTR130* and pBuM-1 satellite DNAs are indicated.

Fig. S4. satDNA consensus sequences from *D. buzzatii, D. seriema* and *D. mojavensis*.

Fig. S5. FISH on polytene chromosomes: **(A)** *CDSTR130* (green) and *pBuM* (red) satDNAs probes on *D. mojavensis,* and **(B)** *CDSTR138* satDNA probe (red) on *D. seriema*.

Supplementary Tables Legends:

Table S1. List of primers used in the satDNA families described in present study.

Table S2. Description of all clusters retrieved from 1834708 reads of *D. buzzatii* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S3. Description of all clusters retrieved from 2144275 reads of *D. seriema* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S4. Description of all clusters retrieved from 2174346 reads of *D. mojavensis* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S5. Description of the ten most abundant clusters of the *D. melanogaster* genome identified by RepeatExplorer. The satDNA families with monomer lengths smaller than 50 bp are highlighted in bold.

Table S6. Main features of 37 *CDSTR198* arrays located on euchromatic regions and their chromosome location according to GenomeBrowser analysis.

Table S7. List of genes associated with *CDSTR198* arrays and their relative positions in relation to *CDSTR198*.

Acknowledgments

We are grateful to Dr. Alfredo Ruiz (Universitat Autònoma de Barcelona) for several insightful discussions during different stages of this work and also for sharing the RNAseq data we used. We also thank Guilherme Borges Dias (Universidade Federal de Minas Gerais) for sequencing *D. seriema*. This work was supported by a grant from "Fundação de Amparo à Pesquisa do Estado de Minas Gerais" (FAPEMIG) (grant number APQ-01563-14) to G.K. LG de Lima was supported with a doctoral fellowship from CAPES. Funding for sequencing was provided by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) - Programa de Excelência Acadêmica (PROEX) - to Programa de Pós Graduação em Genética da UFMG (process CAPES/PROEX 0529/2014). Genomic DNA quality control, library preparation and sequencing were conducted at the Laboratório de Biotecnologia e Marcadores Moleculares of the Universidade Federal de Minas Gerais, with the aid of Dr. Anderson Oliveira do Carmo, Dr. Ana Paula Vimieiro Martins and Dr. Evanguedes Kalapothakis.

References

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, Čech, M. et al., 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, gkw343.

Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K., and Sullivan, B. A, 2016 Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, *26*(10), 1301-1311.

Alkhimova, O. G., Mazurok, N. A., Potapova, T. A., Zakian, S. M., Heslop-Harrison, J. S., and Vershinin, A. V. 2004 Diverse patterns of the tandem repeats organization in rye chromosomes. *Chromosoma*, *113*(1), 42-52.

Baimal, V., Sene, F. M., and Pereira, M. A. O. R. 1983 Heterochromatin and karyotypic differentiation of some neotropical cactus-breeding species of the Drosophila repleta species group. *Genetica*, *60*(2), 81-92.

Barghini, E., Natali, L., Cossu, R. M., Giordani, T., Pindo, M., et al. 2014 The peculiar landscape of repetitive sequences in the olive (Olea europaea L.) genome. *Genome biology and evolution*, *6*(4), 776-791.

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

Beridze, T., 2013.Satellite Dna. Springer Science and Business Media.

Biessmann, H., Zurovcova, M., Yao, J. G., Lozovskaya, E., and Walter, M. F. 2000 A telomeric satellite in Drosophila virilis and its sibling species. *Chromosoma*, *109*(6), 372-380.

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

Blattes, R., Monod, C., Susbielle, G., Cuvier, O., Wu, J., et al. 2006 Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. *The EMBO journal*, *25*(11), 2397-2408.

Bosco, G., Campbell, P., Leiva-Neto, J. T., and 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.

Brajković, J., Feliciello, I., Bruvo-Mađarić, B., and Ugarković, Đ. 2012 Satellite DNA-like elements associated with genes within euchromatin of the beetle Tribolium castaneum. *G3: Genes*| *Genomes*| *Genetics*, *2*(8), 931-941.

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

Casacuberta, E., and Pardue, M. L. 2003 Transposon telomeres are widely distributed in the Drosophila genus: TART elements in the virilis group. *Proceedings of the National Academy of Sciences*, *100*(6), 3363-3368.

Charlesworth, B., Sniegowski, P., and Stephan, L. W. 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, *371*(6494), 215-220.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, et al. 2007 Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203-218.

Cohen, S., and Segal, D. 2009 Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats. *Cytogenetic and genome research*, *124*(3-4), 327-338.

Craddock, E. M., Gall, J. G., and Jonas, M. 2016 Hawaiian Drosophila genomes: size variation and evolutionary expansions. *Genetica*, *144*(1), 107-124.

Dias, G. B., Heringer, P., Svartman, M., and Kuhn, G. C. S. 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.

Dover, G. A., and Tautz, D. 1986 Conservation and divergence in multigene families: alternatives to selection and drift. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *312*(1154), 275-289.

Dover, G. 1982 Molecular drive: a cohesive mode of species evolution. *Nature* 229 (5879): 111-117.

Edgar, R. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.

Feliciello, I., Akrap, I., and Ugarković, Đ. 2015 Satellite DNA modulates gene expression in the beetle Tribolium castaneum after heat stress. *PLoS Genet*, *11*(8), e1005466.

Ferreira, D., Meles, S., Escudeiro, A., Mendes-da-Silva, A., Adega, F., and Chaves, R. 2015. Satellite non-coding RNAs: the emerging players in cells, cellular pathways and cancer. *Chromosome Research*, *23*(3), 479-493.

Fitzgerald, D. J., Dryden, G. L., Bronson, E. C., Williams, J. S., and Anderson, J. N. 1994 Conserved patterns of bending in satellite and nucleosome positioning DNA. *Journal of Biological Chemistry*, *269*(33), 21303-21314.

Franco, F. F., Soto, I. M., Sene, F. M., and Manfrin, M. H. 2008 Phenotypic variation of the aedeagus of Drosophila serido Vilela and Sene (Diptera: Drosophilidae). *Neotropical entomology*, *37*(5), 558-563.

Franco, F. F., Sene, F. M., and Manfrin, M. H. 2008 Molecular characterization of SSS139, a new satellite DNA family in sibling species of the Drosophila buzzatii cluster. *Genetics and Molecular Biology*, *31*(1), 155-159.

Franco, F. F., Silva-Bernardi, E. C. C., Sene, F. M., Hasson, E. R., and Manfrin, M. H. 2010 Intra-and interspecific divergence in the nuclear sequences of the clock gene period in species of the Drosophila buzzatii cluster. *Journal of Zoological Systematics and Evolutionary Research*, *48*(4), 322-331.

Gall, J. G., Cohen, E. H., and Polan, M. L. 1971 Repetitive DNA sequences in Drosophila. *Chromosoma*, *33*(3), 319-344.

Gent, J. I., and Dawe, R. K. 2012 RNA as a structural and regulatory component of the centromere. *Annual review of genetics*, *46*, 443-453.

Gregory, T. R., and Johnston, J. S. 2008 Genome size diversity in the family Drosophilidae. *Heredity*, *101*(3), 228-238.

Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas et al. 2015 Genomics of ecological adaptation in cactophilic Drosophila. *Genome biology and evolution*, *7*(1), 349-366.

Henikoff, S., Ahmad, K., and Malik, H. S. 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, *293*(5532), 1098-1102.

Heslop-Harrison, J. S., Brandes, A., and Schwarzacher, T. 2003 Tandemly repeated DNA sequences and centromeric chromosomal regions of Arabidopsis species. *Chromosome Research*, *11*(3), 241-253.

Jagannathan, M., Warsinger-Pepe, N., Watase, G. J., and Yamashita, Y. M. 2017 Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex. *G3: Genes, Genomes, Genetics*, *7*(2), 693-704.

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

Jurka J. 2012. LTR retrotransposons from fruit fly."; *Repbase Reports* 12(7) 1257-1257.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*(2), 111-120.

Khost, D. E., Eickbush, D. G., and Larracuente, A. M. 2017 Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome Research*.

Kuhn, G. C.S., Ruiz, A., Alves, M. A., and Sene, F. M. 1996 The metaphase and polytene chromosomes of *Drosophila seriema* (repleta group; mulleri subgroup). *Brazilian Journal of Genetics*, *19*, 209-216.

Kuhn, G. C. S., Bollgönn, S., Sperlich, D., and Bachmann, L. 1999 Characterization of a species-specific satellite DNA of *Drosophila buzzatii*. *Journal of Zoological Systematics and Evolutionary Research*, *37*(2), 109-112.

Kuhn, G. C.S., and Sene, F. M. 2005 Evolutionary turnover of two pBuM satellite DNA subfamilies in the *Drosophila buzzatii* species cluster (repleta group): from alpha to alpha/beta arrays. *Gene*, *349*, 77-85.

Kuhn, G. C.S., Franco, F. F., Manfrin, M. H., Moreira-Filho, O., and Sene, F. M. 2007 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 research*, *15*(4), 457-470.

Kuhn, G. C.S., Sene, F. M., Moreira-Filho, O., Schwarzacher, T., and Heslop-Harrison, J. S. 2008 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*, *16*(2), 307-324.

Kuhn, G. C. S., Teo, C. H., Schwarzacher, T., and Heslop-Harrison, J. S. 2009 Evolutionary dynamics and sites of illegitimate recombination revealed in the interspersion and sequence junctions of two nonhomologous satellite DNAs in cactophilic Drosophila species. *Heredity*, *102*(5), 453-464.

Kuhn, G.C,S. and Heslop-Harrison J.S. 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. Cytogenetic and Genome Research 132:79–88

Kuhn, G. C.S., Küttler, H., Moreira-Filho, O., and Heslop-Harrison, J. S. 2012 The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. *Molecular biology and evolution*, 29:7-11.

Langmead, B., and Salzberg, S. L. 2012 Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

Larracuente, A. M. 2014 The organization and evolution of the Responder satellite in species of the Drosophila melanogaster group: dynamic evolution of a target of meiotic drive. *BMC evolutionary biology*, *14*(1), 233.

Leung, W., Shaffer, C. D., Reed, L. K., Smith, S. T., Barshop, W., Dirkes, W., ... and Yuan, H. 2015 Drosophila Muller F elements maintain a distinct set of genomic properties over 40 million years of evolution. *G3: Genes*| *Genomes*| *Genetics*, *5*(5), 719-740.

Lohe, A. R., Hilliker, A. J., and Roberts, P. A. 1993 Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Genetics*, *134*(4), 1149-1174.

Löpez, C. C., Nielsen, L., and Edström, J. E. 1996 Terminal long tandem repeats in chromosomes form Chironomus pallidivittatus. *Molecular and cellular biology*, *16*(7), 3285-3290.

López-Flores, I., and Garrido-Ramos, M. A. 2012 The repetitive DNA content of eukaryotic genomes. In *Repetitive DNA* (Vol. 7, pp. 1-28). Karger Publishers.

Manfrin, M. H., and Sene, F. M. (2006). Cactophilic Drosophila in South America: a model for evolutionary studies. *Genetica*, *126*(1-2), 57-75.

Manfrin, M. H., De Brito, R. O. A., and Sene, F. M. 2001 Systematics and evolution of the Drosophila buzzatii (Diptera: Drosophilidae) cluster using mtDNA. *Annals of the Entomological Society of America*, *94*(3), 333-346.

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novák, P., Schubert, V., et al. 2015 Holocentromeres in Rhynchospora are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. *Proceedings of the National Academy of Sciences*, *112*(44), 13633-13638.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., et al. 2013 Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology*, *14*(1), R10.

Menon, D. U., Coarfa, C., Xiao, W., Gunaratne, P. H., and Meller, V. H. 2014 siRNAs from an X-linked satellite repeat promote X-chromosome recognition in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *111*(46), 16460-16465.

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

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-628.

Mravinac, B., and Plohl, M. 2007 Satellite DNA junctions identify the potential origin of new repetitive elements in the beetle Tribolium madens. *Gene*, *394*(1), 45-52.

Navajas-Pérez, R., Schwarzacher, T., de la Herrán, R., Rejón, C. R., Rejón, M. R., and Garrido-Ramos, M. A. (2006). The origin and evolution of the variability in a Y-specific satellite-DNA of Rumex acetosa and its relatives. *Gene*, *368*, 61-71.

Negre, B., Casillas, S., Suzanne, M., Sánchez-Herrero, E., Akam, M., Nefedov, M., et al. (2005). Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome research*, *15*(5), 692-700.

Nei, M. 1987 *Molecular evolutionary genetics*. Columbia university press.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. 2013 RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792-793.

Ohzeki, J. I., Nakano, M., Okada, T., and Masumoto, H. 2002 CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J Cell Biol*, *159*(5), 765-775.

Oliveira, D. C., Almeida, F. C., O'Grady, P. M., Armella, M. A., DeSalle, R., and Etges, W. J. 2012 Monophyly, divergence times, and evolution of host plant use inferred from a revised

phylogeny of the Drosophila repleta species group. *Molecular Phylogenetics and Evolution*, *64*(3), 533-544.

Pardue, M. L., Rashkova, S., Casacuberta, E., DeBaryshe, P. G., George, J. A., and Traverse, K. L. 2005 Two retrotransposons maintain telomeres in Drosophila. *Chromosome Research*, *13*(5), 443-453.

Pavlek, M., Gelfand, Y., Plohl, M., and Meštrović, N. 2015 Genome-wide analysis of tandem repeats in Tribolium castaneum genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. *Dna research*, *22*(6), 387-401.

Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., and Scheid, O. M. 2010 Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *The Plant Cell*, *22*(9), 3118-3129.

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

Powell, J. R. 1997 *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press.

Rius, N., Guillén, Y., Delprat, A., Kapusta, A., Feschotte, C., and Ruiz, A. 2016 Exploration of the Drosophila buzzatii transposable element content suggests underestimation of repeats in Drosophila genomes. *BMC genomics*, *17*(1), 344.

Romero-Soriano, V., Burlet, N., Vela, D., Fontdevila, A., Vieira, C., and Guerreiro, M. P. G. 2016. Drosophila females undergo genome expansion after interspecific hybridization. *Genome biology and evolution*, *8*(3), 556-561.

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

Rowan, R. G., and Hunt, J. A. 1991. Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for five closely related species of Hawaiian Drosophila. *Molecular biology and evolution*, *8*(1), 49-70.

Ruiz, A., Heed, W. B., and Wasserman, M. 1990. Evolution of the mojavensis cluster of cactophilic Drosophila with descriptions of two new species. *Journal of Heredity*, *81*(1), 30-42.

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., and Camacho, J. P. M. 2016 High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, *6*.

Russo, C. A., Takezaki, N., and Nei, M. 1995 Molecular phylogeny and divergence times of drosophilid species. *Molecular biology and evolution*, *12*(3), 391-404.

Saitou, N., and Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425.

Satović, E., Zeljko, T. V., Luchetti, A., Mantovani, B., and Plohl, M. 2016 Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. *BMC genomics*, *17*(1), 997.

Schwarzacher, T., and Heslop-Harrison, P. 2000 *Practical in situ hybridization*. BIOS Scientific Publishers Ltd.

Sharma, S., and Raina, S. N. 2005 Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. *Cytogenetic and genome research*, *109*(1-3), 15-26.

Shiels, C., Coutelle, C., and Huxley, C. 1997 Contiguous arrays of satellites 1, 3, and β form a 1.5-Mb domain on chromosome 22p. *Genomics*, *44*(1), 35-44.

Silva-Sousa, R., and Casacuberta, E. 2012 Drosophila telomeres: an example of co-evolution with transposable elements. In *Repetitive DNA* (Vol. 7, pp. 46-67). Karger Publishers.

Smit, A. F. A., Hubley, R., and Green, P. 2013 *RepeatMasker Open-4.0. Available from http://www. repeatmasker. org (accessed on 11 February 2016).*

Stephan, W., and Cho, S. 1994 Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics*, *136*(1), 333-341..

Strachan, T., Webb, D., and Dover, G. A. 1985 Transition stages of molecular drive in multiple-copy DNA families in Drosophila. *The EMBO journal*, *4*(7), 1701.

Sun, X., Wahlstrom, J., and Karpen, G. 1997 Molecular structure of a functional Drosophila centromere. *Cell*, *91*(7), 1007-1019.

Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. 2003 Sequence analysis of a functional Drosophila centromere. *Genome research*, *13*(2), 182-194.

Tamura, K., and Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, *10*(3), 512-526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, *28*(10), 2731-2739.

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., and Salzberg, S. L. 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36-46.

Topp, C. N., Zhong, C. X., and Dawe, R. K. 2004 Centromere-encoded RNAs are integral components of the maize kinetochore. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 15986-15991.

Urrego, R., Bernal-Ulloa, S. M., Chavarría, N. A., Herrera-Puerta, E., Lucas-Hahn, A., et al. 2017 Satellite DNA methylation status and expression of selected genes in Bos indicus blastocysts produced in vivo and in vitro. *Zygote*, 1-10.

Villasante, A., Abad, J. P., Planelló, R., Méndez-Lago, M., Celniker, S. E., and de Pablos, B. 2007 Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. *Genome research*, *17*(12), 1909-1918.

Wei, K. H. C., Grenier, J. K., Barbash, D. A., and Clark, A. G. 2014 Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *111*(52), 18793-18798.

Yang, H. P., and Barbash, D. A. 2008 Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome biology*, *9*(2), R39.

Žinić, S. D., Ugarković, D., Cornudella, L., and Plohl, M. 2000 A novel interspersed type of organization of satellite DNAs in Tribolium madens heterochromatin. *Chromosome Research*, *8*(3), 201-212.

Figures:

Figure 1.

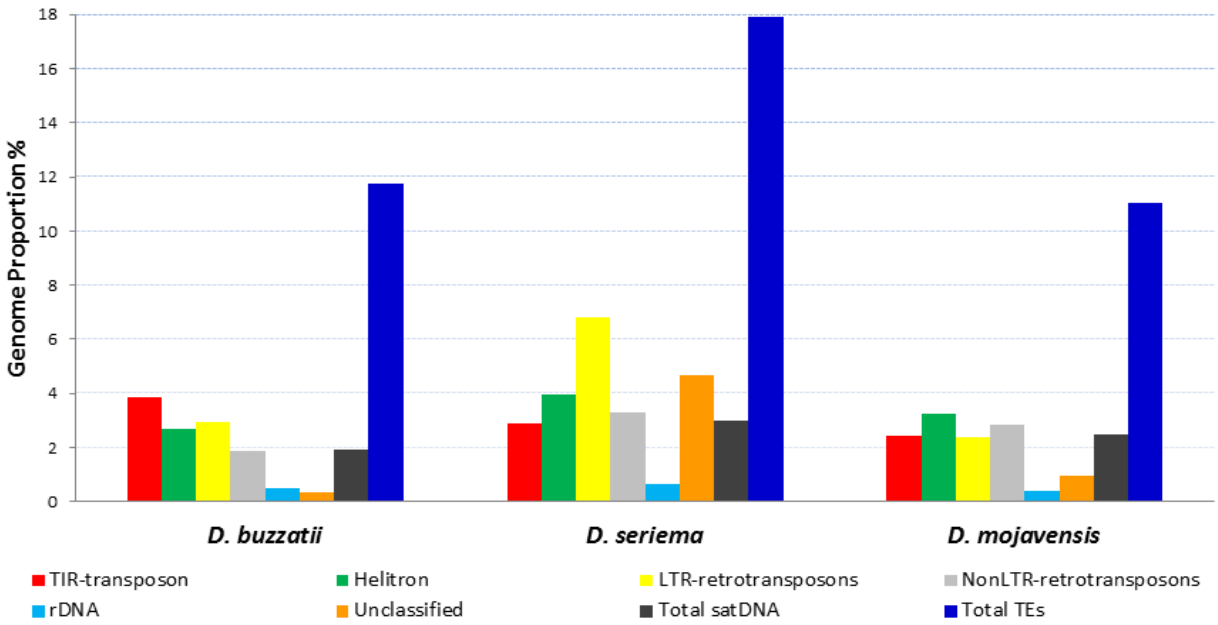

Figure 1. Estimated repetitive DNA abundance in three cactophilic *Drosophila* sepcies.

Figure 2:

Fig. 2: Schematic representation of the BEL3-DM-I transposable element present on RepBase, which is flanked by CDSTR130 satDNA arrays. Blue arrows represent the undescribed 185 bp long terminal repeat of the BEL3-DM element.

Figure 3:

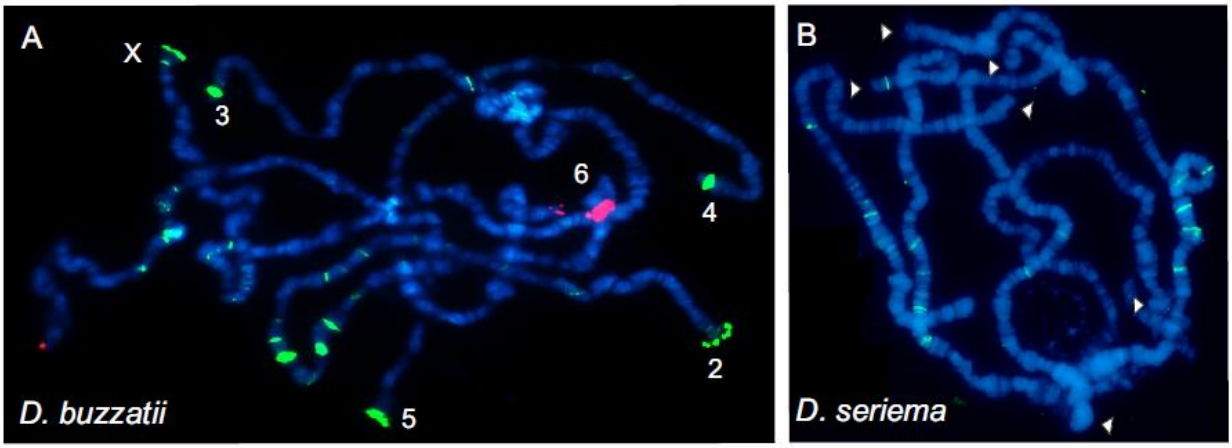

Fig. 3: FISH on polytene chromosomes of *D. buzzatii* (A) and (B) *D. seriema* using satDNA probes for *pBuM* (red) and *CDSTR198* (green) (Arrowheads indicate telomeric regions).

Figure 4.

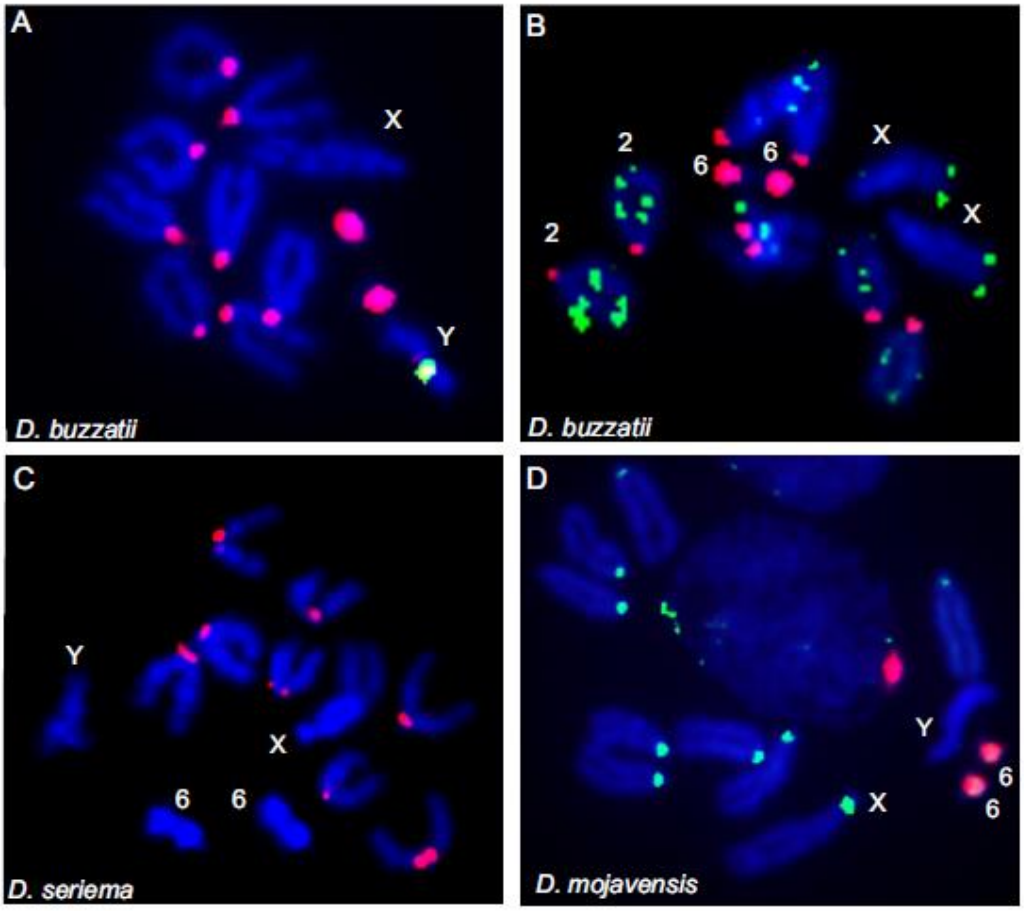

Fig. 4: F**ISH on mitotic chromosomes using satellite DNA probes. (A)** *pBuM-1a* (red) and pBuM-1b (green) satDNA probes on *D. buzzatii;* **B.** *pBuM-1a* (red) and *CDSTR198* (green) probes on *D. buzzatii;* **C.** *CDSTR138* (red) on *D. seriema* (**D)** *CDSTR130* (green) and *pBuM* (red) probes on *D. mojavensis.*

Figure 5

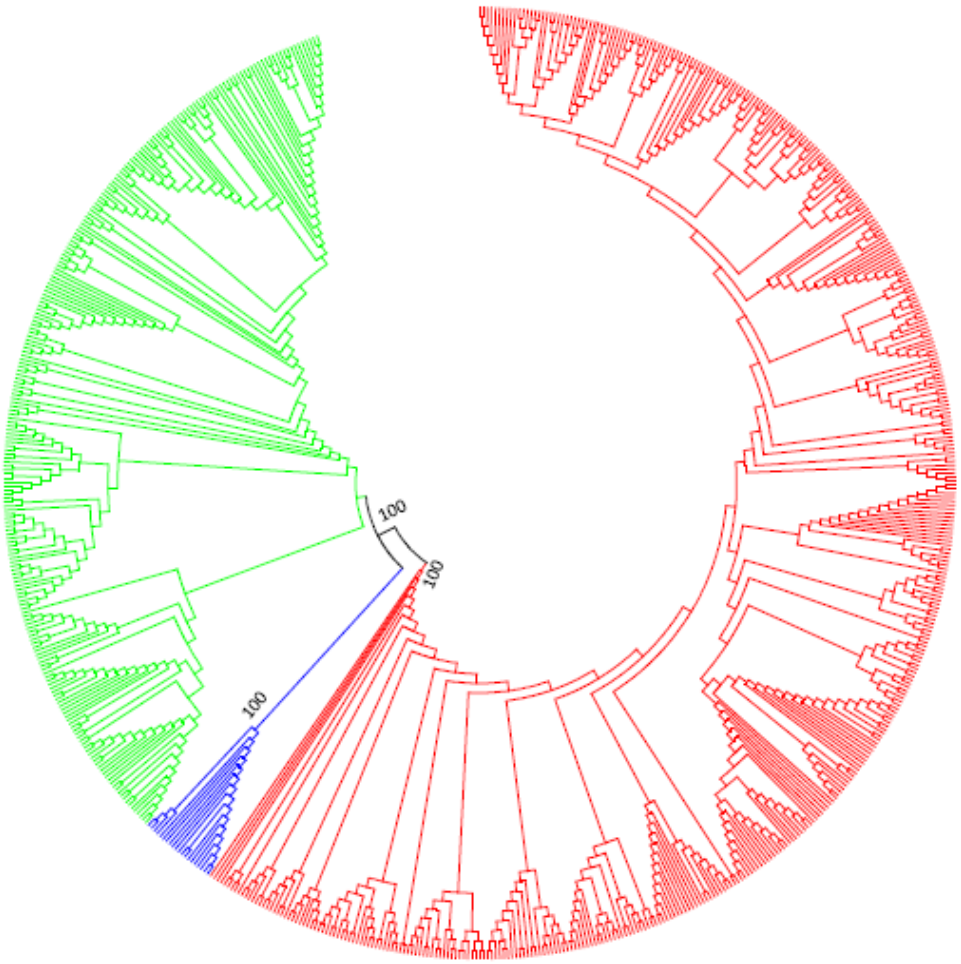

Fig. 5: NJ tree containing a sample of *pBuM* repeats extracted from the sequenced genomes of *Drosophila buzzatii* (green), *D. seriema* (blue) and *D. mojavensis* (red). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

Figure 6:

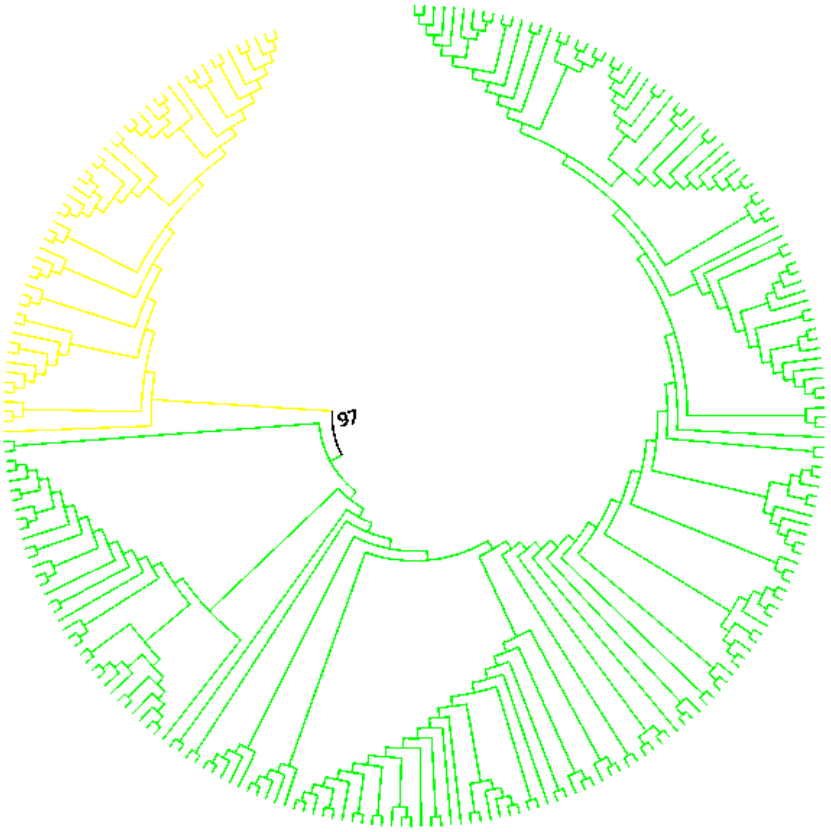

Fig. 6: NJ tree of *pBuM* satDNA repeats retrieved from the *D. buzzatii* assembled genome and previously described on Kuhn et al. (2003). Colored braches evidence Y chromosome specific arrays (yellow) when compared to autosomal arrays (green). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

Figure 7.

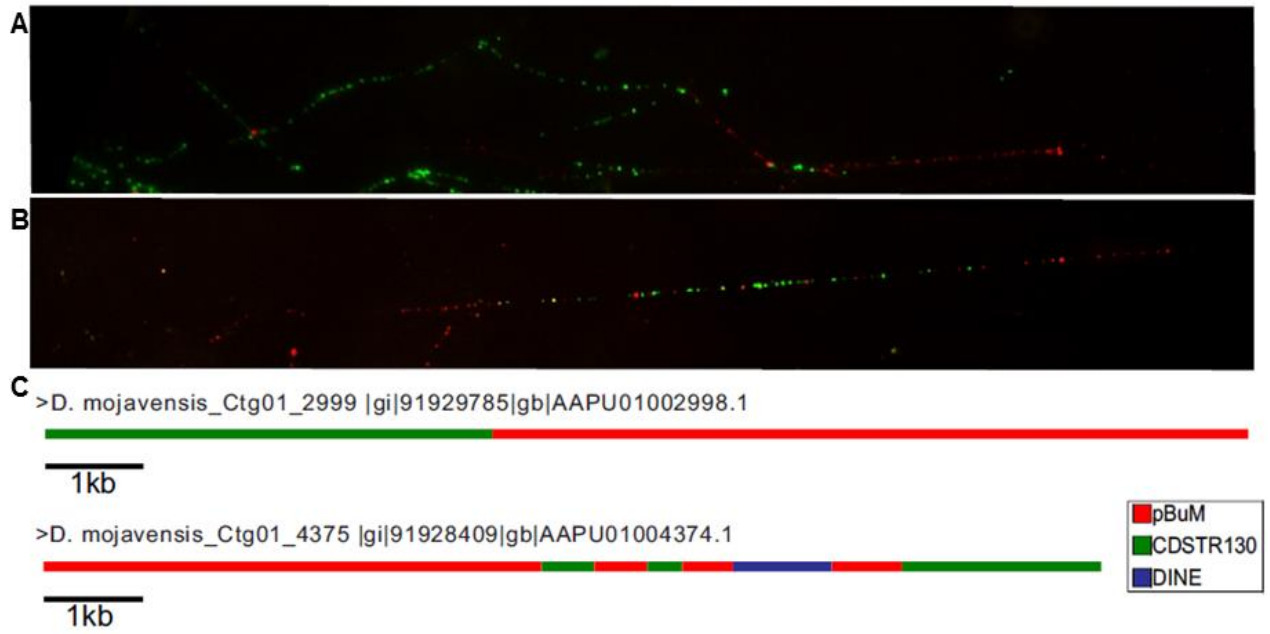

Fig. 7: A-B FISH with *CDSTR*130 (green) and pBuM (red) probes onto extended DNA fibers of *D. mojavensis*. (C) Schematic representation of *CDSTR130* and *pBuM* organization found on *contigs Ctg01_2999(*AAPU01002998.1) and *Ctg01_4375* (AAPU01004374.1 retrieved from the *D. mojavensis* assembled genome.

Figure 8

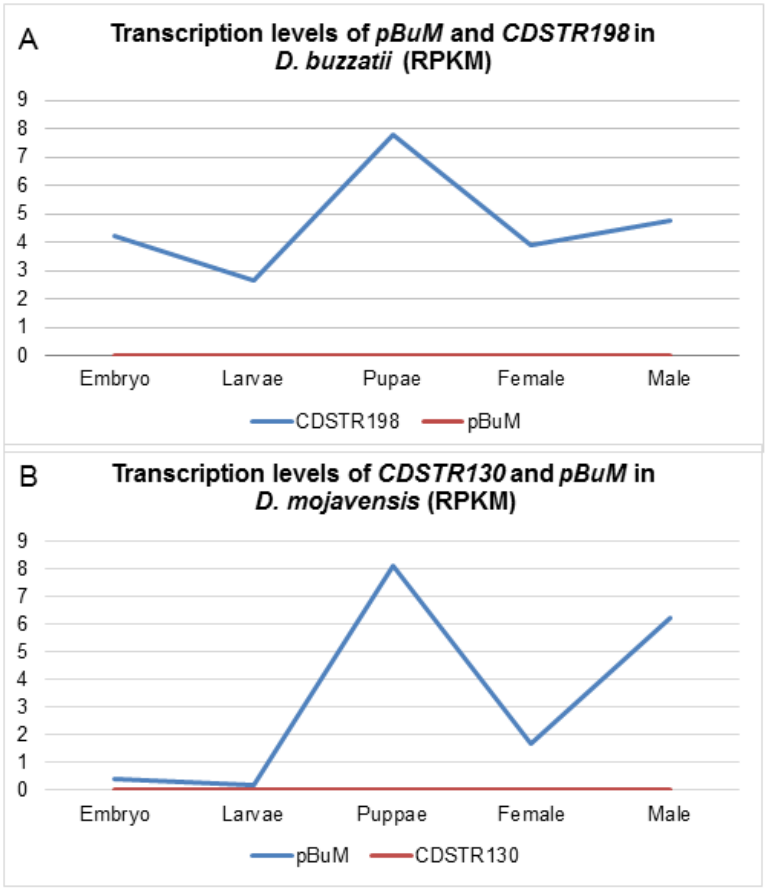

Fig. 8: Transcription profile of satDNA families in *D. buzzatii* (A) and *D. mojavensis* (**B)** on five different developmental stages. Counts were normalized to one million reads.

Figure 9:

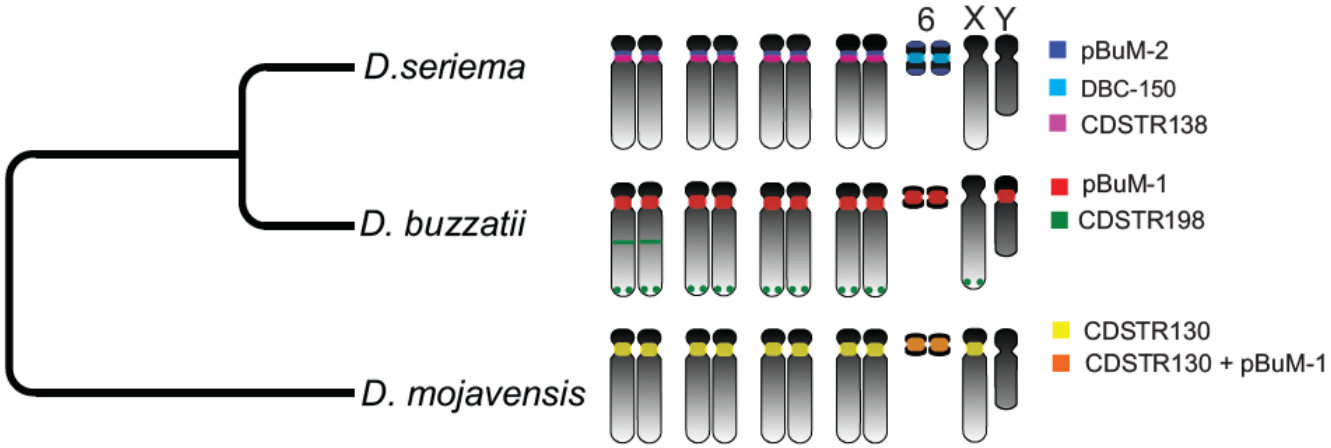

Fig. 9: Representative ideogram showing the chromosomal localization of all satDNAs identified in *D. buzzatii*, *D. seriema* and *D. mojavensis*

List of Tables

Table1. Main features of satellite DNA families present on *D. buzzatii*, *D. seriema* and *D. mojavensis* genomes.

	satDNA family	Monomer Size	GC Content (%)	Copy number (analyzed)	Genomic contribution (%)	Variability (%)
D. buzzatii	*pBuM*	189	29	379	1.71	12.1
	CDSTR198	198	34	79	0.23	13.1
D. seriema	*pBuM-2*	370	23,9	30[a]	1.93	1.9[a]
	DBC-150	150	55.9	5[b]	0.81	11.3[b]
	CDSTR138	138	31.2	386	0.22	12.7
	CDSTR198	198	34.8	67	0.02	15.5
D. mojavensis	*CDSTR130*	130	26.2	929	1.63	13.7
	pBuM	185	26.5	600	0.86	4.1

a Data from Kuhn et al. (2008).
b Data from Kuhn et al. (2007).

Supplementary Material

Supplementary Figures Legends:

Fig. S1. Repetitive clusters (n=122) in *D. buzzatii* identified by RepeatExplorer after clusterization of 270366 reads. Together, these clusters represent 14.7% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-1 and CDSTR198 satellite DNAs are indicated.

Fig. S2. Repetitive clusters (n=328) in *D. seriema* identified by RepeatExplorer after clusterization of 526010 reads. Together, these clusters represent 26.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-2, DBC-150 and *CDSTR138* satellite DNAs are indicated.

Fig. S3. Repetitive clusters (n=217) in *D. mojavensis* identified by RepeatExplorer after clusterization of 323342 reads. Together, these clusters represent 14.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The *CDSTR130* and pBuM-1 satellite DNAs are indicated.

Fig. S4. SatDNA consensus sequences from *D. buzzatii, D. seriema* and *D. mojavensis*.

Fig. S5. FISH on polytene chromosomes: **(A)** *CDSTR130* (green) and *pBuM* (red) satDNAs probes on *D. mojavensis,* and **(B)** *CDSTR138* satDNA probe (red) on *D. seriema*.

Supplementary Tables Legends:

Table S1. List of primers used in the present study.

Table S2. Description of all clusters retrieved from 1834708 reads of *D. buzzatii* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S3. Description of all clusters retrieved from 2144275 reads of *D. seriema* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S4. Description of all clusters retrieved from 2174346 reads of *D. mojavensis* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Table S5. Description of the ten most abundant clusters of the *D. melanogaster* genome identified by RepeatExplorer. The satDNA families with monomer lengths smaller than 50 bp are highlighted in bold.

Table S6. Main features of 37 *CDSTR198* arrays located on euchromatic regions and their chromosome location according to GenomeBrowser analysis.

Table S7. List of genes associated with *CDSTR198* arrays and their relative positions in relation to *CDSTR198*.

Supplementary Figure 1.

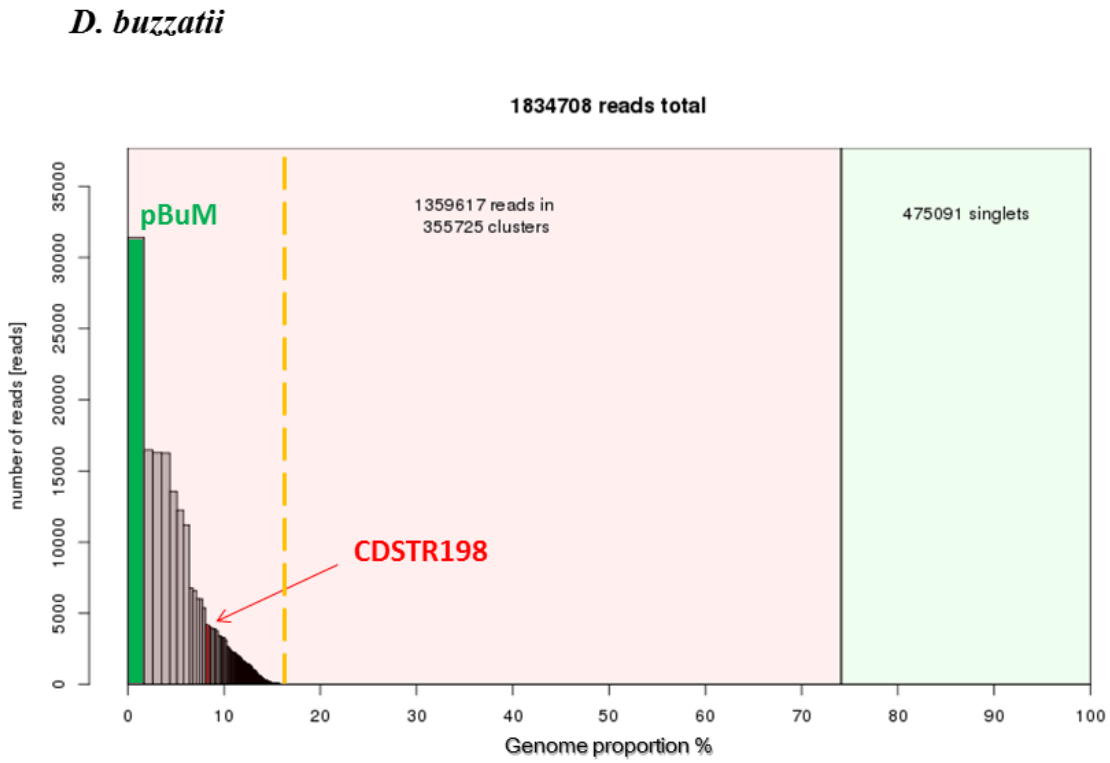

Fig. S1. Repetitive clusters (n=122) in *D. buzzatii* identified by RepeatExplorer after clusterization of 270366 reads. Together, these clusters represent 14.7% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-1 and CDSTR198 satellite DNAs are indicated.

Supplementary Figure 2.

D. seriema

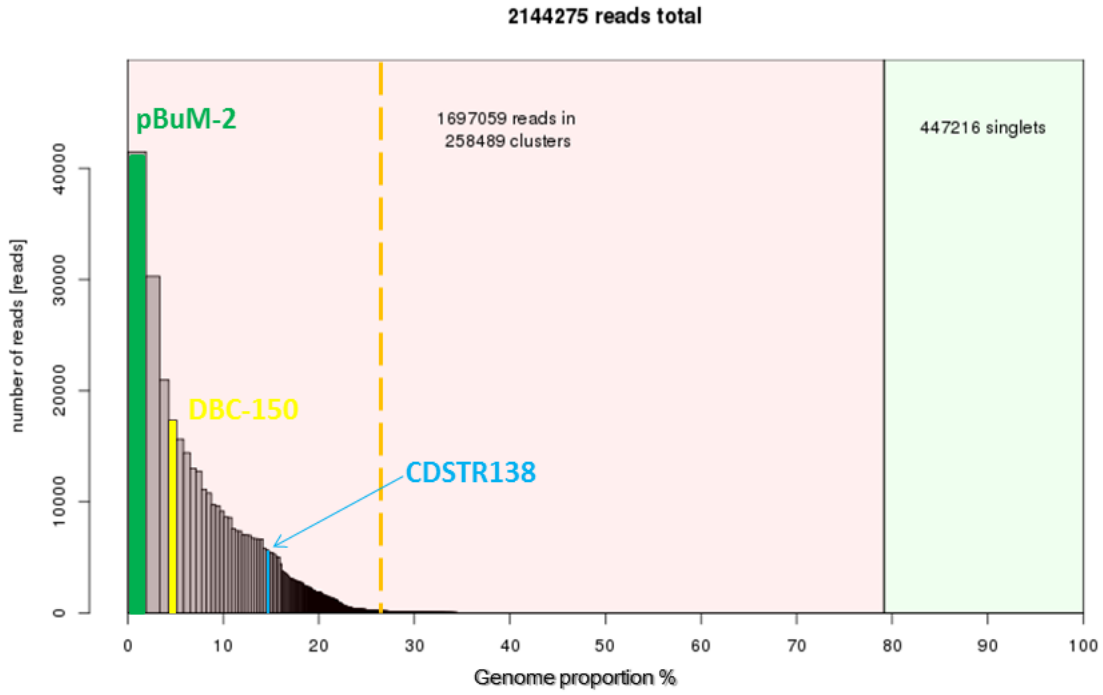

Fig. S2. Repetitive clusters (n=328) in *D. seriema* identified by RepeatExplorer after clusterization of 526010 reads. Together, these clusters represent 26.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The pBuM-2, DBC-150 and *CDSTR138* satellite DNAs are indicated.

Supplementary Figure 3

D. mojavensis

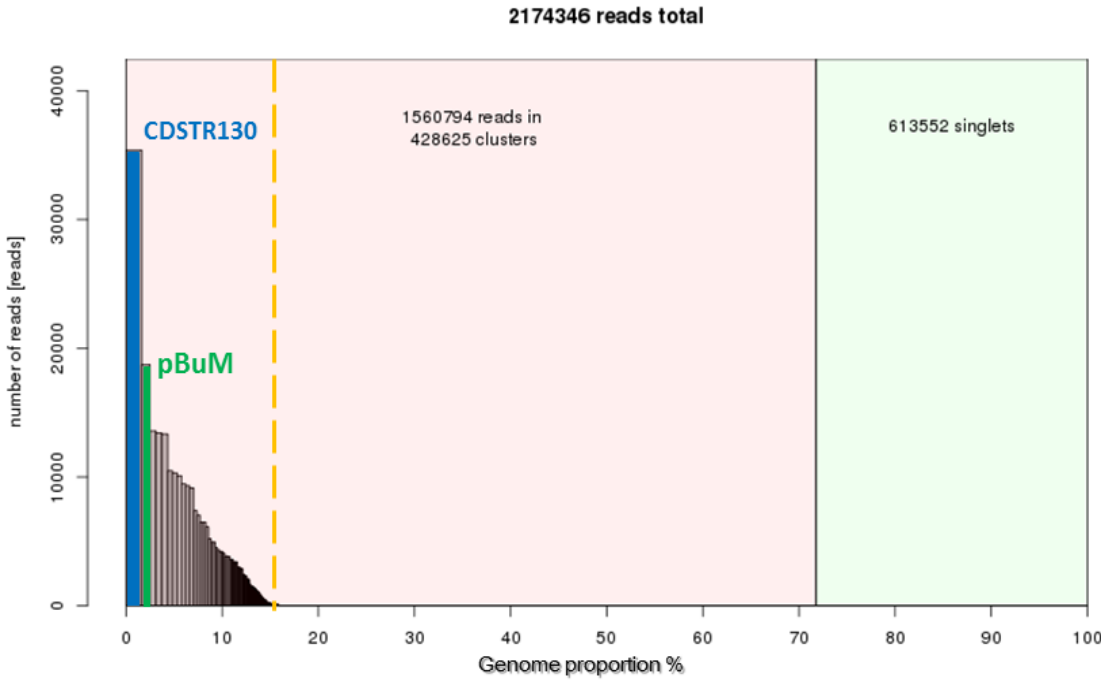

Fig. S3. Repetitive clusters (n=217) in *D. mojavensis* identified by RepeatExplorer after clusterization of 323342 reads. Together, these clusters represent 14.9% of the genome (identified by the yellow traced line). Each bar in the graphic represents a cluster of similar reads. The *CDSTR130* and pBuM-1 satellite DNAs are indicated.

Supplementary Figure 4

>CDSTR130_SatDNA_D.mojavensis_MF170235

CAATAGAGAAAAATTGCAAAATGTGATGGAAAATCAAAAAAAGCAAAGGGATTTTATATTTTAAAG
CTTTGATAACAAAATATTGAAAGATATTCGGCAATAAACTGGTAAATATTG

>CDSTR138_SatDNA_D.seriema_MF170236

CAACTGTTGATTTTTGTATATAGAATACGAATAAAATCAAATATATGGTGTAGAAGAGTATATTTG
CGCCGATTTACCATGAACATAACACTAGTTTGATTTTTTATTCATTTATAGTGAGTATAGGCATAT
CTGGGC

>CDSTR198_SatDNA_D.buzzatii_MF170237

AAGGTAGAAAGGTAGTTGGTGAGATAAACCAGAAAAAGAGCTAAAAACGGCTAAAAACGGCTAGAA
AATAGCCAGAAAGGTAGATTGAACATTAATGGGCAAATGGATGGATAAATAAGACTGGTCATCATC
CAATGAACAGAATCATGATTAAGAGATAGAAATATGATTAGAAAGTAGGATAGAAAGGTTAGAAAG

>pBuM-1a_SatDNA_D.buzzatii_MF170238

GCAAAAGACTCCGTCAATTAGAAAACAAAAAATGTTATAGTTTTGAGGATTAACCGGCAAAAACCG
TATTATTTGTCATTTGATTTCTTTATGGAATACCGTTTTAGAAGCGTCTTTTATCGTATTACTCAG
ATATATCTTAAGATTTAGCATAATCTAAGAACTTTTTGAAATATTCACATTTGTCCA

>pBuM-1b_SatDNA_D.buzzatii_Y_MF170239

TAAAATTACTACTTGAAACTAGAAAGAAAGAAAAGTTATAGTTTTGAGGTTTAACCGGCAAAAATC
GTATTATTTATCATTAGATTTCTTTATAGCATGCCGTTTATAAGCGTGTCTTATCGGATTATTCAG
ATATATTGCAAAATTTAACATAGCTCGAGACCTTTTTGAAATATTAACATTAATCCA

>pBuM-1_SatDNA_D.mojavensis_MF170240

CCTGAAATGCGGTATAAGATCAGAAAATTGACTTTTTTCTATGTATAACAGGCAATAACAGGACTA
TTTCGGGGCCGATTTCAGTACGACTTTCTTTGTTGGAAGCATCTTTTAGAGCCCTATCTATTGACG
TATTCAGATTTTCAATGGGCCAACAACTTTTGAAATATTTACATTATTATCCC

Fig. S4. satDNA consensus sequences from *D. buzzatii, D. seriema* and *D. mojavensis*.

Supplementary Figure 5.

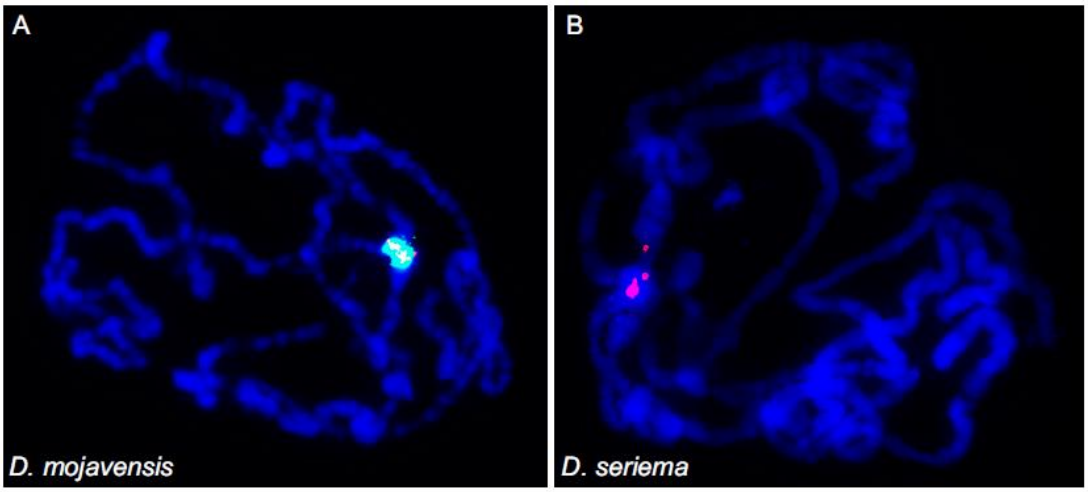

Fig. S5. FISH on polytene chromosomes: **(A)** *CDSTR130* (green) and *pBuM* (red) satDNAs probes on *D. mojavensis,* and **(B)** *CDSTR138* satDNA probe (red) on *D. seriema*.

Supplementary Table 1. List of primers used in the present study.

	Primer sets	Primer sequences
D. buzzatii	*pBuM-a*	Fwd: CAATTAGAAAACAGGAAATTG
		Rev: GACGGAGTCTTTTGCTGGACAA
	pBuM-1b	Fwd: GAAACTAGAAGGAAATAAACG
		Rev: CAAGGAGTCCTTTTATGGATTA
	CDSTR198	Fwd: GGTAGATTGAACACTAATGGGC
		Rev: CCTTTCTGCCTTTTTTCTAGCCG
D. seriema	*CDSTR138*	Fwd:CCGATTTTCCATGAACATAACC
		Rev: CGCAAATATACTCTACTACACC
D. mojavensis	*CDSTR130*	Fwd: CAATAGAGGCAAATGAATTTAG
		Rev: CATTACTAACCAGTTTATTGCCG
	pBuM-1a	Fwd: GCCGATTTCAGTACGACTTACTTCG
		Rev: CCAGAAATAGTCCTGTTATTGCCTG

Table S2. Description of all clusters retrieved from 1834708 reads of *D. buzzatii* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Cluster	*Genome Proportion %*	*RepeatMasker Annotation*	*Reads number on each cluster*	*Outside reads with similarity %*
CL1	**1.710**	***pBuM-1***	**31409**	**0.00038**
CL2	0.898	RC.Helitron	16475	0.19870
CL3	0.888	DNA.hAT.Pegasus	16285	2.70400
CL4	0.886	DNA.hAT.hobo	16263	3.69400
CL5	0.739	Helitron	13567	4.80500
CL6	0.668	RC.Helitron	12252	8.96000
CL7	0.611	Non-LTR CR1	11208	0.11510
CL8	0.370	Galileo	6790	0.55450
CL9	0.360	Homo6	6601	6.44000
CL10	0.328	LINE-like	6024	0.03731
CL11	0.327	Galileo	6007	0.49990
CL12	0.294	rRNA	5393	0.05911
CL13	**0.229**	***CDSTR198***	**4202**	**0.00650**
CL14	0.226	Unclassified	4142	0.08527
CL15	0.217	LTR BEL	3982	0.06588
CL16	0.215	PERI	3940	2.02800
CL17	0.213	LTR BEL	3914	0.30090
CL18	0.205	Galileo/Helitron	3766	6.62800
CL19	0.188	rRNA	3442	0.16500
CL20	0.185	LTRGypsy	3387	0.08860
CL21	0.181	Non-LTR_R1	3317	0.05204
CL22	0.179	LTR Gypsy/Minicopia	3290	0.05014
CL23	0.170	LTR Gypsy	3122	0.03179
CL24	0.145	LTR BEL	2653	0.04116
CL25	0.143	LTR Gypsy	2622	0.42640
CL26	0.136	LTR Osvaldo	2493	0.12570
CL27	0.129	PARISa	2364	0.08243
CL28	0.125	INVADER6_I	2292	0.14160
CL29	0.124	LTR BEL	2279	0.42260
CL30	0.121	Galileo	2214	0.29560
CL31	0.120	LTR Copia	2198	0.28850
CL32	0.113	LTR Gypsy	2070	0.13140
CL33	0.110	RC.Helitron	2025	0.28540
CL34	0.109	LTR Osvaldo	1996	0.06629
CL35	0.104	Unclassified	1917	0.27710
CL36	0.100	Unclassified	1835	0.05137
CL37	0.095	Non-LTR Jockey	1736	0.02687
CL38	0.089	LTR BEL	1640	0.57820
CL39	0.089	Non-LTR R1	1632	0.21130
CL40	0.087	LTRBEL	1592	0.02742

CL41	0.084	Homo9	1537	0.29000
CL42	0.082	LTR Gypsy	1501	0.03540
CL43	0.079	Non-LTR Jockey	1451	0.06051
CL44	0.078	LTR Copia	1437	0.00347
CL45	0.078	Non-LTR Jockey	1428	0.02450
CL46	0.077	LTR Copia	1411	0.11840
CL47	0.074	Non-LTR_I	1351	0.07546
CL48	0.073	DNA.TcMar.Tc1	1340	0.52960
CL49	0.069	Non-LTR Jockey	1260	0.38600
CL50	0.061	Unclassified	1126	0.12680
CL51	0.059	Non-LTR Jockey	1076	0.23030
CL52	0.058	LTR BEL	1066	0.03978
CL53	0.056	LTR Copia	1019	0.15520
CL54	0.055	DNA.CMC.Transib	1015	0.49800
CL55	0.054	LTR Gypsy	987	0.00000
CL56	0.053	Non-LTR RTE.BovB	982	0.00000
CL57	0.050	LTR Gypsy	909	0.12640
CL58	0.049	DNA.TcMar.Tc1	891	0.20170
CL59	0.045	Non-LTR Jockey	828	0.03882
CL60	0.043	DNA.hAT-9	795	0.06751
CL61	0.038	Unclassified	705	0.68000
CL62	0.038	RNA	699	0.00454
CL63	0.037	RC.Helitron	687	0.09555
CL64	0.035	LTR Copia	644	0.00000
CL65	0.035	DNA.hAT.hobo	639	0.30280
CL66	0.035	LTR Copia	636	0.07575
CL67	0.034	LTR Gypsy	627	0.10790
CL68	0.034	Unclassified	625	0.19720
CL69	0.033	Galileo	604	0.25630
CL70	0.030	DNA.TcMar.Tc1	560	0.07143
CL71	0.030	RC.Helitron	543	0.03627
CL72	0.029	Unclassified	530	0.00000
CL73	0.028	DNA.TcMar.Tc1	509	0.22400
CL74	0.026	Non-LTR Jockey	480	0.47450
CL75	0.026	Unclassified	477	0.19390
CL76	0.025	LTR Gypsy	450	0.00000
CL77	0.023	LTR Copia	430	0.00000
CL78	0.023	LTR Gypsy	429	0.36040
CL79	0.022	DNA.hAT.Pegasus	412	0.08403
CL80	0.022	Rep2-DF	396	0.13160
CL81	0.021	LINE jockey	383	0.00000
CL82	0.021	Unclassified	382	0.00000
CL83	0.021	DNA.TcMar.Tc1	376	0.27770
CL84	0.020	DNA.hAT.Ac	375	0.05963
CL85	0.020	LTR Copia	368	0.04246
CL86	0.019	LTR Gypsy	344	0.28300
CL87	0.018	DNA.TcMar.Tc1	335	0.38710

CL88	0.018	LTRGypsy	333	0.05133
CL89	0.018	LTR Copia	332	0.07948
CL90	0.017	LTR Gypsy	314	0.33650
CL91	0.017	LTR Gypsy	314	1.91400
CL92	0.017	DNA.PiggyBac	312	0.00000
CL93	0.017	LTR BEL	309	0.00000
CL94	0.017	DNA.hAT.hobo	309	0.05537
CL95	0.017	LTR Copia	307	0.20480
CL96	0.016	Non-LTR Jockey	299	0.32750
CL97	0.016	LTR Gypsy	298	0.15710
CL98	0.016	DNA.CMC.Transib	297	1.41100
CL99	0.016	Non-LTR Jockey	293	0.00000
CL100	0.016	DNA.TcMar.Tc1	291	0.47050
CL101	0.016	LTR Gypsy	289	0.00000
CL102	0.015	Unclassified	268	0.00000
CL103	0.015	Unclassified	267	0.25610
CL104	0.014	Unclassified	259	0.00000
CL105	0.014	LTR Gypsy	249	0.07072
CL106	0.013	LTR Gypsy	245	0.09872
CL107	0.013	Unclassified	244	0.23550
CL108	0.013	Unclassified	241	0.11280
CL109	0.013	LTR Gypsy	235	0.41250
CL110	0.013	LTR BEL	231	0.00000
CL111	0.012	LTR Gypsy	223	0.16910
CL112	0.012	DNA.TcMar.	221	0.20410
CL113	0.012	DNA.TcMar.	220	0.81220
CL114	0.012	DNA.TcMar.Tc1	213	0.72610
CL115	0.011	LTR Gypsy	205	0.18590
CL116	0.011	Unclassified	204	0.00000
CL117	0.011	DNA Sola1-type	199	0.21910
CL118	0.011	Unclassified	195	0.00000
CL119	0.011	DNA.CMC.Transib	194	2.55300
CL120	0.010	LTR Gypsy	190	0.26460
CL121	0.010	DNA.CMC.Transib	189	0.10960
CL122	0.010	LTR Gypsy	186	0.00000

Table S3. Description of all clusters retrieved from 2144275 reads of *D. seriema* by Repeat Explorer. The satDNA families analyzed in this study are highlighted in bold red.

Cluster	***Genome Proportion%***	***RepeatMasker Annotation***	***Reads number on each cluster***	***Outside reads with similarities %***
CL1	**1.930**	**pBuM-2**	**41472**	**0.176**
CL2	1.410	RC.Helitron	30287	24.010
CL3	0.978	Homo6	20966	32.230
CL4	**0.809**	**DBC-150**	**17349**	**0.784**
CL5	0.730	RC.Helitron	15646	26.060

CL6	0.673	Non-LTR CR1	14427	5.296
CL7	0.607	PERI	13026	24.800
CL8	0.595	RC.Helitron	12761	42.540
CL9	0.519	LTR Gypsy	11132	2.677
CL10	0.504	SSS139	10798	19.010
CL11	0.455	Non-LTR R1	9753	0.749
CL12	0.450	Non-LTR R1	9650	1.907
CL13	0.428	LTR BEL	9170	2.345
CL14	0.403	rRNA	8643	0.104
CL15	0.401	Galileo	8605	6.392
CL16	0.355	LTR Gypsy	7615	2.232
CL17	0.345	Unclassified	7408	0.310
CL18	0.344	Unclassified	7385	0.163
CL19	0.329	RC.Helitron	7044	31.960
CL20	0.328	Galileo	7041	9.260
CL21	0.326	Non-LTR Jockey	6993	2.331
CL22	0.314	Non-LTR Jockey	6743	4.968
CL23	0.313	LTR BEL	6705	1.253
CL24	0.311	Unclassified	6674	1.049
CL25	0.311	LTR BEL	6658	6.008
CL26	0.272	LTR Gypsy	5826	1.442
CL27	**0.266**	***CDSTR138***	**5709**	**0.893**
CL28	0.257	Non-LTR R1	5502	3.017
CL29	0.253	LTR Gypsy	5433	4.013
CL30	0.249	LTR Gypsy	5343	1.029
CL31	0.238	rRNA	5094	1.924
CL32	0.234	LTR Copia	5022	3.146
CL33	0.206	rRNA	4421	3.936
CL34	0.176	LTR Gypsy	3781	1.878
CL35	0.171	LTR BEL	3663	2.266
CL36	0.167	LTR Gypsy	3583	5.610
CL37	0.160	Unclassified	3438	2.967
CL38	0.152	LTR Gypsy	3257	2.395
CL39	0.147	LTR Gypsy	3143	6.713
CL40	0.144	Unclassified	3082	4.121
CL41	0.143	LTR Copia	3076	2.503
CL42	0.140	LTR BEL	3011	1.328
CL43	0.137	LTR Gypsy	2933	0.852
CL44	0.136	RC.Helitron	2910	0.687
CL45	0.133	LTR Gypsy	2856	2.066
CL46	0.132	Unclassified	2830	0.177
CL47	0.129	LTR Gypsy	2766	3.868
CL48	0.129	Unclassified	2759	8.191
CL49	0.123	LTR Gypsy	2647	2.909

CL50	0.115	RC.Helitron/DNA.hAT.Pegasus	2474	0.000
CL51	0.115	Non-LTR I	2473	3.275
CL52	0.114	LTR BEL	2440	3.238
CL53	0.111	LTR Gypsy	2378	0.925
CL54	0.111	DNA.Ginger	2373	6.153
CL55	0.109	Unclassified	2343	0.427
CL56	0.107	LTR Gypsy	2289	3.102
CL57	0.100	LTR Gypsy	2152	2.138
CL58	0.098	Non-LTR Jockey	2100	0.048
CL59	0.096	Galileo	2061	2.232
CL60	0.094	LTR BEL	2023	3.114
CL61	0.091	LTR Gypsy	1960	2.092
CL62	0.090	LTR Gypsy	1922	1.769
CL63	0.089	RC.Helitron	1919	1.928
CL64	0.089	LTR Gypsy	1910	1.571
CL65	0.089	LTR Gypsy	1899	2.791
CL66	0.088	LTR Gypsy	1895	3.219
CL67	0.088	Non-LTR Jockey	1883	1.009
CL68	0.085	LTR Gypsy	1832	4.039
CL69	0.082	Non-LTR Jockey	1758	2.446
CL70	0.079	Non-LTR Jockey	1698	4.005
CL71	0.078	DNA.TcMar.Tc1	1665	3.363
CL72	0.075	LTR Gypsy	1612	1.365
CL73	0.074	Unclassified	1596	0.125
CL74	0.074	Unclassified	1586	0.568
CL75	0.072	Unclassified	1555	0.000
CL76	0.072	DNA.CMC.Transib	1545	24.140
CL77	0.071	LTR BEL	1530	1.046
CL78	0.069	LTR Gypsy	1486	2.826
CL79	0.067	DNA.hAT.Tip100	1435	3.902
CL80	0.067	DNA.TcMar.Tc1	1429	2.799
CL81	0.066	LTR ERVL	1411	0.071
CL82	0.066	Unclassified	1406	0.854
CL83	0.065	Non-LTR Jockey	1400	0.643
CL84	0.064	DNA.TcMar.Tc1	1366	2.928
CL85	0.062	LTR Copia	1340	0.298
CL86	0.061	Galileo	1311	3.051
CL87	0.061	LTR Copia	1306	2.374
CL88	0.059	Non-LTR Jockey	1263	3.959
CL89	0.059	Non-LTR Jockey	1261	1.983
CL90	0.057	LTR Gypsy	1215	2.881
CL91	0.056	DNA.hAT.Ac	1200	5.917

CL92	0.054	Unclassified	1165	0.000
CL93	0.054	DNA.TcMar.Tc1	1163	4.729
CL94	0.049	Non-LTR RTE	1043	1.534
CL95	0.048	LTR Gypsy	1019	0.687
CL96	0.045	LTR Gypsy	969	0.516
CL97	0.044	LTR Gypsy	953	3.043
CL98	0.044	Unclassified	940	0.319
CL99	0.044	LTR Gypsy	934	0.107
CL100	0.043	Unclassified	932	0.000
CL101	0.042	LTR Gypsy	903	0.111
CL102	0.042	DNA.Maverick	895	0.335
CL103	0.041	Unclassified	869	0.115
CL104	0.038	LTR Gypsy	810	2.469
CL105	0.037	DNA.hAT.Pegasus	800	4.625
CL106	0.037	Unclassified	785	0.255
CL107	0.035	Unclassified	749	0.000
CL108	0.033	LTR Gypsy	715	2.378
CL109	0.033	Unclassified	709	0.000
CL110	0.033	Non-LTR R1	709	1.975
CL111	0.033	LTR Copia	701	0.856
CL112	0.032	Unclassified	688	0.000
CL113	0.030	Unclassified	639	0.157
CL114	0.029	Unclassified	631	0.000
CL115	0.029	LTR Gypsy	625	3.520
CL116	0.029	LTR Copia	624	3.526
CL117	0.029	LTR Gypsy	623	3.371
CL118	0.029	Unclassified	620	0.000
CL119	0.028	Unclassified	605	0.000
CL120	0.028	LTR Gypsy	600	1.167
CL121	0.028	Non-LTR I	595	0.840
CL122	0.027	Unclassified	585	0.000
CL123	0.027	Unclassified	579	0.000
CL124	0.027	DNA.PIF	575	0.000
CL125	0.026	Unclassified	550	0.000
CL126	0.026	Unclassified	549	0.000
CL127	0.025	LTR Gypsy	541	0.000
CL128	0.025	Galileo	532	4.511
CL129	0.025	DNA.TcMar.Tc1	531	4.520
CL130	0.025	LTR Copia	528	0.379
CL131	0.024	RC.Helitron	521	0.000
CL132	0.024	LTR Gypsy	518	1.158
CL133	**0.024**	**CDSTR198**	**513**	**0.195**
CL134	0.024	LTR Gypsy	503	0.000
CL135	0.023	LTR Copia	499	1.804
CL136	0.023	Unclassified	493	0.000

CL137	0.022	Unclassified	483	0.207
CL138	0.022	Unclassified	480	0.000
CL139	0.022	Unclassified	479	0.000
CL140	0.022	Galileo	476	4.202
CL141	0.022	Unclassified	470	0.000
CL142	0.022	Unclassified	467	0.214
CL143	0.021	DNA.TcMar.Tc1	449	3.786
CL144	0.020	Unclassified	438	0.000
CL145	0.020	Unclassified	436	0.000
CL146	0.020	Unclassified	436	0.000
CL147	0.020	LTR Gypsy	434	1.152
CL148	0.020	Unclassified	432	0.232
CL149	0.020	RC.Helitron	430	3.953
CL150	0.020	Non-LTR L1	429	0.000
CL151	0.020	LTR Gypsy	426	0.000
CL152	0.020	Unclassified	426	0.000
CL153	0.020	Unclassified	424	0.236
CL154	0.020	Unclassified	423	0.000
CL155	0.020	Unclassified	420	0.000
CL156	0.020	Unclassified	420	0.000
CL157	0.020	Unclassified	420	0.000
CL158	0.019	Unclassified	419	0.000
CL159	0.019	Non-LTR Jockey	412	1.456
CL160	0.019	RC.Helitron	410	2.439
CL161	0.019	DNA.hAT.hATm	410	0.000
CL162	0.019	DNA.hAT.hobo	409	1.711
CL163	0.019	Unclassified	407	0.000
CL164	0.019	Unclassified	405	0.000
CL165	0.019	Unclassified	404	0.000
CL166	0.019	Unclassified	399	0.000
CL167	0.018	LTR BEL	397	0.000
CL168	0.018	LTR Gypsy	396	0.505
CL169	0.018	Unclassified	394	0.000
CL170	0.018	LTR BEL	394	1.523
CL171	0.018	Unclassified	392	0.000
CL172	0.018	Unclassified	389	0.000
CL173	0.018	Unclassified	387	0.000
CL174	0.018	DNA.CMC.Transib	387	8.010
CL175	0.018	Unclassified	383	0.000
CL176	0.018	Unclassified	381	0.000
CL177	0.018	Unclassified	380	0.000
CL178	0.018	Unclassified	379	0.000
CL179	0.018	Unclassified	379	0.000

CL180	0.017	DNA.TcMar.Tc1	374	1.337
CL181	0.017	Unclassified	373	0.000
CL182	0.017	Unclassified	371	0.000
CL183	0.017	Unclassified	370	0.000
CL184	0.017	Unclassified	369	0.000
CL185	0.017	DNA.TcMar.Tc1	365	1.370
CL186	0.017	Unclassified	365	0.000
CL187	0.017	Unclassified	365	0.000
CL188	0.017	Unclassified	365	0.000
CL189	0.017	DNA.hAT	365	0.000
CL190	0.017	DNA.CMC.Transib	364	0.000
CL191	0.017	Unclassified	364	0.000
CL192	0.017	Unclassified	364	0.000
CL193	0.017	Unclassified	363	0.000
CL194	0.017	Unclassified	358	0.000
CL195	0.017	LTR Copia	356	0.281
CL196	0.017	Unclassified	356	0.000
CL197	0.017	Unclassified	354	0.000
CL198	0.017	Unclassified	353	0.000
CL199	0.017	Unclassified	353	0.000
CL200	0.016	Unclassified	349	0.286
CL201	0.016	rDNA	349	0.000
CL202	0.016	tRNA	349	0.000
CL203	0.016	Unclassified	349	0.000
CL204	0.016	Unclassified	348	0.000
CL205	0.016	Unclassified	347	0.000
CL206	0.016	Unclassified	346	0.000
CL207	0.016	Unclassified	344	0.000
CL208	0.016	LTR BEL	343	0.000
CL209	0.016	Unclassified	339	0.000
CL210	0.016	Unclassified	337	0.000
CL211	0.016	Unclassified	335	0.000
CL212	0.015	Unclassified	332	0.000
CL213	0.015	Unclassified	331	0.000
CL214	0.015	Unclassified	328	0.000
CL215	0.015	Unclassified	328	0.000
CL216	0.015	Unclassified	326	0.000
CL217	0.015	LTR ERVK	326	0.000
CL218	0.015	Unclassified	324	0.000
CL219	0.015	Unclassified	323	0.000
CL220	0.015	Unclassified	323	0.000
CL221	0.015	Unclassified	319	0.000
CL222	0.015	Unclassified	313	0.000
CL223	0.015	Unclassified	312	0.000
CL224	0.015	Unclassified	311	0.000

CL225	0.015	Unclassified	311	0.000
CL226	0.015	Unclassified	311	0.000
CL227	0.014	Unclassified	308	0.000
CL228	0.014	Unclassified	304	0.000
CL229	0.014	DNA.CMC.Chapaev	304	0.000
CL230	0.014	Unclassified	302	0.000
CL231	0.014	Unclassified	300	0.000
CL232	0.014	Unclassified	300	0.000
CL233	0.014	Unclassified	299	0.000
CL234	0.014	LTR Gypsy	299	0.000
CL235	0.014	Unclassified	299	0.000
CL236	0.014	Unclassified	298	0.000
CL237	0.014	Unclassified	296	0.000
CL238	0.014	Non-LTR CR1	296	0.000
CL239	0.014	DNA.CMC.EnSpm	295	0.000
CL240	0.014	Unclassified	295	0.000
CL241	0.014	Unclassified	292	0.000
CL242	0.014	Unclassified	291	0.000
CL243	0.013	Unclassified	289	0.000
CL244	0.013	Non-LTR I	287	0.000
CL245	0.013	Unclassified	287	0.000
CL246	0.013	Unclassified	286	0.000
CL247	0.013	Unclassified	286	0.000
CL248	0.013	Unclassified	283	0.000
CL249	0.013	DNA.TcMar.Tc	282	2.482
CL250	0.013	Unclassified	279	0.000
CL251	0.013	LTR Copia	277	0.000
CL252	0.013	Unclassified	276	0.000
CL253	0.013	Unclassified	274	0.000
CL254	0.013	Unclassified	274	0.000
CL255	0.013	Unclassified	273	0.000
CL256	0.013	Unclassified	273	0.000
CL257	0.013	Unclassified	273	0.000
CL258	0.013	Unclassified	273	0.000
CL259	0.013	DNA.Maverick	271	1.107
CL260	0.013	Unclassified	268	0.000
CL261	0.013	DNA.PiggyBac	268	0.746
CL262	0.013	Unclassified	267	0.000
CL263	0.012	Unclassified	265	0.000
CL264	0.012	Unclassified	264	0.000
CL265	0.012	Unclassified	263	0.000
CL266	0.012	Unclassified	263	0.000
CL267	0.012	Unclassified	262	0.000

CL268	0.012	Unclassified	261	0.000
CL269	0.012	Unclassified	259	0.000
CL270	0.012	Unclassified	259	0.000
CL271	0.012	Unclassified	257	0.000
CL272	0.012	LTR Gypsy	257	0.000
CL273	0.012	Unclassified	255	0.000
CL274	0.012	Unclassified	254	0.000
CL275	0.012	LTR ERVK	254	0.000
CL276	0.012	Unclassified	253	0.000
CL277	0.012	Unclassified	251	0.000
CL278	0.012	Unclassified	251	0.000
CL279	0.012	Unclassified	250	0.000
CL280	0.012	Unclassified	250	0.000
CL281	0.012	Unclassified	248	0.000
CL282	0.011	Unclassified	247	0.000
CL283	0.011	Unclassified	246	0.000
CL284	0.011	Unclassified	245	0.000
CL285	0.011	Unclassified	245	0.000
CL286	0.011	Unclassified	244	0.000
CL287	0.011	Unclassified	243	0.000
CL288	0.011	Unclassified	243	0.411
CL289	0.011	Unclassified	241	0.000
CL290	0.011	Unclassified	240	0.000
CL291	0.011	Unclassified	238	0.000
CL292	0.011	Unclassified	238	0.000
CL293	0.011	LTR Copia	238	0.840
CL294	0.011	Unclassified	237	0.000
CL295	0.011	Unclassified	237	0.000
CL296	0.011	Unclassified	236	0.000
CL297	0.011	Unclassified	235	0.000
CL298	0.011	Unclassified	235	0.000
CL299	0.011	Unclassified	234	0.000
CL300	0.011	Unclassified	234	0.000
CL301	0.011	Unclassified	234	0.000
CL302	0.011	Unclassified	233	0.000
CL303	0.011	Unclassified	233	0.000
CL304	0.011	Unclassified	231	0.000
CL305	0.011	Unclassified	230	0.000
CL306	0.011	Unclassified	229	0.000
CL307	0.011	Unclassified	227	0.000
CL308	0.011	DNA.PIF.ISL2EU	226	0.000
CL309	0.011	Unclassified	226	0.000
CL310	0.011	Unclassified	225	0.000
CL311	0.011	Unclassified	225	0.000
CL312	0.010	Unclassified	224	0.000

CL313	0.010	Unclassified	224	0.000
CL314	0.010	Non-LTR Penelope	224	0.000
CL315	0.010	Unclassified	223	0.000
CL316	0.010	Unclassified	222	0.000
CL317	0.010	LTR Gypsy	221	0.000
CL318	0.010	Unclassified	221	0.000
CL319	0.010	Unclassified	220	0.000
CL320	0.010	Unclassified	219	0.000
CL321	0.010	Non-LTR CR1	219	0.000
CL322	0.010	Unclassified	219	0.000
CL323	0.010	Unclassified	218	0.000
CL324	0.010	Unclassified	218	0.000
CL325	0.010	Unclassified	218	0.459
CL326	0.010	Unclassified	218	0.000
CL327	0.010	Unclassified	216	0.000
CL328	0.010	Unclassified	214	0.000

Table S4. Description of all clusters retrieved from 2174346 reads of *D. mojavensis* by RepeatExplorer. The satDNA families analyzed in this study are highlighted in bold red.

Cluster	***Genome Proportion %***	***RepeatMasker Annotation***	***Reads number on each cluster***	***Outside reads with similarity %***
CL1	**1.630**	***CDSTR130***	**35395**	**0.494**
CL2	**0.861**	***pBuM***	**18724**	**0.417**
CL3	0.625	RC.Helitron	13585	41.430
CL4	0.616	Homo6	13402	45.750
CL5	0.614	RC.Helitron	13345	42.750
CL6	0.482	Homo6	10483	33.900
CL7	0.475	RC.Helitron	10323	42.020
CL8	0.464	Non-LTR Jockey	10096	0.366
CL9	0.435	RC.Helitron	9461	21.380
CL10	0.429	Galileo	9320	2.157
CL11	0.420	Non-LTR R1	9132	1.818
CL12	0.340	RC.Helitron	7382	31.140
CL13	0.323	Unclassified	7027	0.896
CL14	0.299	RC.Helitron	6496	3.079
CL15	0.298	Non-LTR R1	6480	0.972

CL16	0.281	LTR BEL	6119	0.082
CL17	0.239	Unclassified	5196	0.019
CL18	0.227	Non-LTR R1	4944	3.621
CL19	0.227	Non-LTR CR1	4935	2.411
CL20	0.208	Non-LTR Jockey	4515	0.975
CL21	0.200	rRNA	4355	0.987
CL22	0.194	LTR BEL	4216	0.427
CL23	0.194	Non-LTR R1	4208	0.048
CL24	0.188	PERI	4096	29.050
CL25	0.178	Non-LTR Jockey	3860	0.130
CL26	0.177	LTR Gypsy	3856	0.778
CL27	0.175	Non-LTR Jockey	3814	0.000
CL28	0.166	Non-LTR Jockey /TART	3605	3.523
CL29	0.166	rRNA	3601	1.000
CL30	0.157	LTR Gypsy	3420	1.637
CL31	0.157	LTR Gypsy	3409	0.264
CL32	0.156	LTR BEL	3392	0.884
CL33	0.138	LTR Gypsy	3007	0.200
CL34	0.137	LTR Gypsy	2978	0.000
CL35	0.135	RC.Helitron	2925	9.983
CL36	0.132	LTR Gypsy	2867	1.604
CL37	0.111	RC.Helitron	2413	2.445
CL38	0.110	DNA.hAT.Ac	2399	0.125
CL39	0.107	Rep2-DF	2337	2.696
CL40	0.106	LTR Copia	2302	0.174
CL41	0.097	DNA.TcMar.Tc1	2103	3.281
CL42	0.094	LTR BEL	2045	1.125
CL43	0.093	DNA.TcMar.Tc1	2032	1.280
CL44	0.077	DNA.TcMar.Tc1	1681	0.833
CL45	0.072	LTR BEL	1570	3.694
CL46	0.072	Unclassified	1559	1.668
CL47	0.071	DNA.CMC.Transib	1540	0.779
CL48	0.069	LTR Gypsy	1492	0.335
CL49	0.067	LTR Gypsy	1448	0.691
CL50	0.063	Unclassified	1379	2.321
CL51	0.062	DNA.hAT.Tip100	1360	2.426
CL52	0.062	Unclassified	1358	0.074
CL53	0.059	Unclassified	1284	0.000
CL54	0.054	LTR Gypsy	1177	1.359
CL55	0.053	Non-LTR R1	1164	0.172
CL56	0.052	Non-LTR Jockey	1126	0.178
CL57	0.051	Unclassifeid	1105	0.000
CL58	0.050	LTR BEL	1093	0.366
CL59	0.049	Galileo	1066	1.876

CL60	0.048	LTR BEL	1045	0.766
CL61	0.041	Non-LTR Jockey	883	1.133
CL62	0.040	LTR Gypsy	870	0.575
CL63	0.040	LTR Gypsy	869	0.000
CL64	0.037	DNA.TcMar.Mariner	804	0.746
CL65	0.036	DNA.TcMar.Tc1	780	0.000
CL66	0.034	DNA.TA-1	731	1.231
CL67	0.033	DNA-1_RPr	724	0.000
CL68	0.030	LTR Gypsy	661	1.210
CL69	0.030	RC.Helitron	657	0.457
CL70	0.029	Non-LTR Jockey	627	0.478
CL71	0.026	LTR BEL	565	1.593
CL72	0.026	LTR Gypsy	562	0.000
CL73	0.025	Unclassified	537	0.000
CL74	0.024	LTR Gypsy	527	0.000
CL75	0.024	Non-LTR Jockey	518	1.158
CL76	0.023	DNA.TcMar.Tc1	502	0.398
CL77	0.023	LTR ERV1	496	0.202
CL78	0.022	DNA.TcMar.Tc1	482	5.394
CL79	0.022	DNA.TcMar.Tc1	472	0.424
CL80	0.021	rDNA	454	0.000
CL81	0.020	LTR BEL	442	0.226
CL82	0.020	DNA.piggyBac	441	1.134
CL83	0.020	DNA.TcMar.	438	1.598
CL84	0.019	Non-LTR Jockey	417	0.000
CL85	0.018	Non-LTR CR1	381	7.612
CL86	0.017	Non-LTR Jockey	367	0.273
CL87	0.017	Non-LTR Jockey	362	0.276
CL88	0.016	LTR Gypsy	354	1.412
CL89	0.014	Unclassifeid	310	18.710
CL90	0.014	LINE	298	0.000
CL91	0.013	DNA.hAT.Ac	293	0.341
CL92	0.013	DNA.hAT.Pegasus	289	2.422
CL93	0.013	DNA.hAT.hobo	288	3.819
CL94	0.013	LTR Gypsy	282	0.000
CL95	0.013	DNA.TcMar.Tc1	278	2.158
CL96	0.013	LTR Gypsy	273	0.000
CL97	0.012	Unclassifeid	256	0.000
CL98	0.012	Unclassifeid	255	0.000
CL99	0.011	Unclassifeid	248	0.000
CL100	0.011	Non-LTR Jockey	244	0.410
CL101	0.011	LTR BEL	243	0.000
CL102	0.011	DNA.PIF.Harbinger	242	1.653

CL103	0.011	DNA.TcMar.Tc1	237	1.266
CL104	0.011	LTR Gypsy	235	0.425
CL105	0.010	DNA.CMC.Transib	226	0.443
CL106	0.010	DNA.hAT.hobo	224	3.125
CL107	0.010	DNA.hAT.Pegasus	217	3.687
CL108	0.010	Unclassified	217	0.000
CL109	0.010	DNA.TcMar.Tc1	217	0.461

Table S5. Description of the ten most abundant clusters of the *D. melanogaster* genome identified by RepeatExplorer. The satDNA families with monomer lengths smaller than 50 bp are highlighted in bold.

Cluster	Sequence Annotation	Monomer size (bp)	Genome proportion %	Number of reads on each cluster
1	**TGTTATTCTA**	**10**	**1.750**	**86909**
2	**AGAGA**	**5**	**0.715**	**35548**
3	*1.688*	360	0.551	27393
4	**AGACA**	**5**	**0.484**	**24099**
5	*1.688*	360	0.453	22558
6	*1.688*	360	0.437	21759
7	***dodeca***	**11-12**	**0.380**	**18912**
8	*1.688*	360	0.340	16906
9	*Responder*	240	0.323	16044
10	LTR-Gypsy	-	0.293	14577

Table S6. Main features of 37 *CDSTR198* arrays located on euchromatic regions and their chromosome location according to GenomeBrowser analysis.

Contig	***Array Size (bp)***	***Copy number***	***Scaffold***	***Chromosome***
713	300	1,51	1	2
714	521	2,63	1	2
834	643	3,24	1	2
894	563	2,84	1_2	2
1013	585	2,95	1_2	2
1263	908	4,58	2	2
1281	462	2,33	2	2
1732	557	2,81	1_2	2
2138	348	1,75	6	3
2239	912	4,6	6	3
2290	814	4,1	6	3
3200	1044	5,27	11	4
4011	884	4,46	13	5
4217	627	3,16	14	5
4994	243	1,22	22	4

5012	530	2,67	22	4
5199	608	3,07	18_1	3
5294	450	2,27	18_1	3
5693	741	3,74	18_2	3
6645	233	1,17	28	X
6973	804	4,06	22	4
8355	287	1,45	40	4
8688	625	3,15	41	5
8952	405	2,04	31	4
9897	499	2,52	52	2
12072	243	1,22	6	3
12476	659	3,33	83	3
13070	678	3,42	24	3
13095	680	3,43	24	3
13237	652	3,29	68	5
14801	521	2,63	121	5
15065	692	3,49	123	4
16146	616	3,11	174	Unmapped
18240	404	2,04	111	5
20000	607	3,06	18	3
27251	322	1,62	2	2
31812	196	0,99	2	2

Table S7. List of genes associated with *CDSTR198* arrays and their relative positions in relation to *CDSTR198*.

Contig	*Array Size (bp)*	*Position of CDSTR198 relative to gene*	*Distance of CDSTR198 array to gene*	*Gene annotation*	*D. melanogaster orthologous*	*Scaffold/ chromosome*
714	**521**	Intron	-	EVM prediction scaffold1.763	FBpp0082178	1/2
894	**563**	Intron	-	EVM prediction scaffold 1.1077	FBpp0111784	1/2
1013	**585**	Intron	-	EVM prediction scaffold1.1420	FBpp0303067	1/2
2138	**348**	Intron	-	EVM prediction scaffold6.61	FBpp0080953	6/3
4011	**884**	5’	1.4 kb	EVM prediction scaffold13.141	FBpp0167799	13/5
4217	**627**	5’	750 bp	FBpp0170282	-	14/5
5012	**530**	5’	500 bp	EVM prediction scaffold22.170	-	22/4
5199	**608**	5’	4 kb	Gene.g4503.tl	-	18/3
6645	**233**	5’	3 kb	EVM prediction scaffold28.31	FBpp0073879	28/X
6973	**804**	5’	800 bp	EVM prediction scaffold22.28	FBpp0072761	22/4
8355	**287**	5’	1.9 kb	EVM prediction scaffold40.3	FBpp0072462	40/4
8688	**625**	Intron	-	EVM prediction scaffold41.61	FBpp0112268	41/5
8952	**405**	5’	500 bp	EVM prediction scaffold31.206	FBpp0079016	31/4
13070	**678**	5’	800 bp	EVM prediction scaffold24_1.4	-	24/3
13095	**680**	5’	300 bp	SNAPPosition:scaffold24 _preditec gene_268763..270446	-*	24/3
14801	**521**	5’	3,9 kb	EVM prediction scaffold121.6	FBpp0074830	121/5

*This predicted gene has no orthologous on *D. melanogaster* or *D. mojavensis* according to *Drosophila buzzatii* Genome Project (http://dbuz.uab.cat; Guillen et al. 2015)

6. Capítulo III – Artigo a ser submetido à revista PlosOne: In Depth Satellitome Analysis of 36 Drosophila Genomes Reveals Sources of Genome Size Variation

In Depth Satellitome Analysis of 36 *Drosophila* Genomes Reveals Sources of Genome Size Variation

Leonardo G. de Lima[1]*, Gustavo C.S. Kuhn[1]

[1]Universidade Federal de Minas Gerais, Laboratório de Citogenômica Evolutiva, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Avenida Presidente Antônio Carlos, 6627 – Pampulha, 31270-901. Belo Horizonte, Brazil.

*Corresponding author: E-mail: leonardogdlima@gmail.com

Telephone: 5531-34092612

FAX: 5531-34092567

Abstract

Satellite DNAs (satDNA) are ubiquitously present in eukaryotic genomes and have been recently associated with several biological roles. However, in *Drosophila* only a few studies focused on satDNA sequences in a phylogenetic context. We addressed this question by conducting a satellitome analysis in 36 species from genus *Drosophila* and measuring the correlation of these sequences with genome size evolution using *RepeatExplorer* pipeline to identify *de novo* satDNAs. Herein we described 172 satDNA families, being 133 of them newly described satDNA sequences. Repeat analysis within a phylogenetic framework has revealed the profound divergent nature of satDNA sequences in *Drosophila* genomes. We observed that the satDNA content varied from 0.54% of *D. arizonae* genome to 27.4% of *D. montana.* Moreover, monomers size and GC% also showed extremes variations, from 2-570 bp and 9.1-71.4%, respectively. We also described the maintenance of satDNA families shared among closely related species, confirming the satDNA library hypothesis. Notably, we found that the *1.688* satDNA family is present in a wide range of species which diverged ~27 Mya. Finally, we found that changes in genome size of *Sophophora* are positively correlated with transposable element abundance, whereas in *Drosophila* subgenus satDNA sequences strongly correlate with genome size variation. This indicates that in both subgenera genome size evolution seems to occur through modulation of different repetitive elements. Here, we present the most comprehensive satellitome analysis in *Drosophila* species and the largest satDNA sequence catalog to date, which could aid future discoveries either on satDNA evolution as on genome assembly.

Keywords: *Satellite DNA*; Repetitive DNA; *RepeatExplorer*.

Introduction

Satellite DNA (satDNA) sequences are present in virtually all eukaryotic genomes studied to date and are usually associated with heterochromatin (Charlesworth 1994). SatDNAs are generally formed by long tandem arrays in which the monomers (motifs) are repeated in a head-to-tail fashion and can represent up to 50% of the genome content in some species (reviewed by Plohl 2012; Schmidt and Heslop-Harrison 1998). Initially, the low gene content of the heterochromatin led to the misconception that satDNAs have no essential function, and thus, these sequences were traditionally considered as junk DNA (Ohno 1972). However, it is now clear that satDNA sequences can play a role in several cellular processes such as kinetochore assembly, X chromosome recognition and meiotic chromosome segregation (Kuhn 2015).

Despite their importance for genome organization, function and evolution, satDNAs have been only rarely studied in *Drosophila* species with sequenced genomes and in eukaryotes in general. The study of satDNAs and its use as cytogenetic marker usually requires *a priori* knowledge of its sequences. In *Drosophila,* satDNA investigations began relatively later when compared to other animal, but with time proceeded very intensely (Laird and McCarthy 1968; Gall et al. 1971). The initial satDNAs studies on *Drosophila* used cesium chloride density gradient to characterize the most abundant sequences in *D. virilis*, *D. melanogaster* and *D. hydei* genomes (Gall et al. 1974; Barnes et al. 1978; Renkawitz 1979). In the past decades, satDNAs have been mostly studied from a small sample of cloned repeats obtained by biased experimental approaches (usually by restriction digestion and/or PCR), isolated from one or few species. (Brutlag et al. 1977; Waring and Pollack 1987; Bonaccorsi and Lohe 1991; Bachmann and Sperlich 1993; Kuhn et al. 1999, 2008).

SatDNAs abundance can differ dramatically between species, apparently evolving by array expansion and shrinkage of related repeat variants (Nijman and Lestra 2001; Slamovits et al. 2001). In several *Drosophila* species, satDNAs account for more than 30% of the genome and amplification/contraction events of distinct satDNA families have been identified

as an important factor in shaping the architecture and size of the *Drosophila* genome (Bosco et al. 2007). Moreover, speciation process may be associated with satDNA given that rapid changes in copy number can trigger rapid genome changes (Gregory and Johnston 2008; Ferree and Barbash 2009).

Due to its repetitive nature, satDNAs create ambiguities in the processes of aligning and assembling Next-Generation Sequencing (NGS) data (Treangen and Salzberg 2012). As result, most of the recent genome reports, if not all, lack a curated characterization of satDNAs and use *RepeatMasker* database to identify the overall presence of repetitive families. Indeed, despite the large repertoire of sequenced genomes available among *Drosophila* species, up to now only 23 satDNAs families from *Drosophila* are deposited in the *Genbank* or *Repbase* databases (accessed June 2017). Moreover, the majority of recent studies focused on deciphering the satDNA dynamics in *D. melanogaster* subgroup or *D. virilis* subgroup (Kuhn et al. 2012; Dias et al. 2014; Gallach 2014; Larracuente 2014; Garavis et al. 2015; Khost et al. 2017; De Lima et al. *in preparation*).

Despite the effort to describe satDNA sequences in *Drosophila*, studies focused on specific evolutionary questions are scarce and done on a hardly representative number of species and species groups (reviewed by Plohl 2012). Thus, a detailed identification of all satDNA families is a major step to understand their evolutionary links and how they can influence genome evolution. The recent availability of the *RepeatExplorer* pipeline (Novak et al. 2013) allowed us to access the whole collection of repetitive sequences from a given genome using large NGS datasets, fostering the characterization of new satDNAs (Ruiz-Ruano et al. 2016; Palacios-Gimenez et al. 2017; Araújo et al. 2017; De Lima et al. 2017). Additionally, recent sequencing efforts and computational approaches on *D. melanogaster* genome have suggested that the *1.688* satDNA family is a crucial component of centromeric functions, chromosome missegregation in hybrids, dosage compensation and heterochromatin formation (Cattani and Presgraves 2012; Ferree and Prasad 2012; Menon

et al. 2014; Rosic et al. 2014). These data have strengthened the argument that satDNA is an important mechanism factor for genomic organization and species adaptation.

In this paper, we performed a high-throughput analysis of the satellitome from 36 *Drosophila* species which revealed 133 new satDNA sequences present on *Drosophila* genomes and confirmed the presence of the previously described satDNA sequences. Moreover, we also compared the maintenance of several satDNA families throughout *Drosophila* phylogeny, as well as the burst of several species-specific satDNA families. Finally, we discuss the implications of these findings for the study of genome size evolution. Altogether, our work represents the most comprehensive genome comparison focusing satDNA evolution on *Drosophila* and includes a large satDNA database that will improve *Drosophila* genome annotation on a greater scale.

Material and Methods

Genomic data

The Illumina sequence reads from the 36 *Drosophila* species were downloaded directly from the EBI Short Read Archive (SRA) publicly available genome sequences. All SRA files used in the present work are described in the Supplementary Table 1. Moreover, genome size estimations used in the present study were retrieved from Animal Genome Size Database (www.genomesize.com), Bosco et al. (2007) and Hjelmen and Johnston (2017).

Identification of Satellite DNA sequences

Repetitive elements were identified using the *RepeatExplorer* Galaxy server (http://www.*RepeatExplorer*.org/; Novák et al. 2013). *RepeatExplorer* identifies repetitive elements *de novo* using a graph-based method to group reads into discrete clusters based on all-by-all blast similarity. This analysis is unbiased and uses a large repertoire of sequences, resulting in a higher quality analysis and with a larger variability of sequences due its overall sequence clusterization. All reads used in this analysis were trimmed at 100

bp and removed sequencing adapters, excepted for *D. persimilis* reads that were generated with only 76 bp. Moreover, the cut-off quality values were set at 30 and only reads with 90% the percent of bases with values higher than the quality cut-off were kept. The clusterization threshold was explicitly set to 90% sequence similarity spanning at least 65% of the read length. Only clusters with genomic proportion equal or higher to 0.01% were analyzed in this study. The analysis with values bellow this threshold is not recommended by *RepeatExplorer* developers (Novák et al. 2013). Posteriorly, reads that shared high sequence similarity were clustered and further aligned and partially assembled to each cluster using CAP3 (parameters: -O -p 80 -o 40; Huang and Madan 1999; Novák et al. 2010). All clusters annotated were masked on *RepeatMasker* (Smit et al. 2013, 2014, 2015) using the Metazoa database of the *RepeatMasker* during annotation. This approach provides information about repeat quantities (estimated from the number of reads that comprise each cluster), the relationship among clusters and outputs from BLASTn and BLASTx (Altschul et al. 1990) similarities searches on *Repbase* database.

Moreover, only clusters consisting of at least 20 reads were annotated. This cut-off was low enough to fully capture highly abundant repeats in all species, yet remaining computationally tractable. Whenever a significant number of reads from two distinct clusters match the similarity parameters, *RepeatExplorer* indicates these clusters as ‘connected component’, pointing to a potential relationship between the repeats.

All clusters were manually curated using visual inspection of the *RepeatExplorer* assembled contigs, BLAST searches at NCBI nt/nr database and Tandem Repeats Finder to identify satDNA sequences (default parameters – max. motif size 2000 nucleotides; Benson 1999). However, uncharacterized sequences are absent in the databases. Thus, to describe all tandemly arranged sequences in this study was necessary to confirm its period size and its nucleotide structure. This manual annotation was crucial to the reliability of the satDNA description. Accordingly, we used Dot-plot analysis to retrieve this information without the previous description of satDNA sequence identity. These plots were generated with the

Dotlet applet (Junier and Pagni 2000) with a 15 bp word size and 60% similarity cutoff. Posterior to the tandem repeats identification step, we ran a second round of RepeatMasker searches using *Drosophila* database to characterize possible transposable element (TE) internal repeats. Manual curation was also required to exclude mitochondrial and microbial contaminants.

Moreover, to confirm the presence of tandemly repeated arrays on the *Drosophila* species genomes, all satDNA sequences consensus identified by *RepeatExplorer* methods were used as query on BLASTn searches on assembled genome data on NCBI whole genome sequences (WGS) database. To examine patterns on a broader scale, we also mined each satDNA sequence and performed alignment analyses after a random generator (www.random.org) to select 20 contigs retrieved from BLAST output with e-value lower than 10^{-5}.

Alignment and phylogenetic tree reconstruction

For each satDNA family, the nucleotide sequences corresponding to a pool of each reference element were aligned using MUSCLE (Edgar 2004) in the MEGA7 platform (Kumar et al. 2016). We added some previously described sequences from *Drosophila* satDNAs to confirm the sequence initial point of satDNA families present on other organisms available in Genbank (Table S2). We determined the nucleotide evolution model to be used in the phylogenetic reconstructions using Jmodeltest2 (Darrida et al. 2012). This analysis allowed us to reveal the same evolutionary model T93 to best explain our data for each family. Tree reconstructions were performed by the Neighbor-joining algorithm implemented in MEGA 7 (Kumar et al. 2016) with 1000 bootstrap replicates using the respective evolutionary model.

Statistical Analyses

To infer the correlation between satDNA, repetitive components and genome sizes of *Drosophila* species, we used the Spearman's correlation coefficient (ρ). This correlation test was used due the nonparametric nature of genome size and repetitive DNA evolution. Spearman's correlation ρ was calculated by $\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$, where i is paired score.

Results and Discussion

A bioinformatic *de novo* identification reveals a highly divergent *Drosophila* satellitome

Herein, we identified 172 satDNA-like sequences of *Drosophila*, including representatives of several more species groups. The use of the *RepeatExplorer* pipeline in this study aimed to minimize the underrepresentation of previous estimations of satDNAs and other repetitive families done on flies by means of genome assemblies (*Drosophila* 12 genomes Consortium 2007; Guillén et al. 2015). The overall *de novo* clusterization of reads using at least 0.35-fold of the genome coverage (Table S1) suggests that most, if not all, repetitive sequences were clustered and analyzed in this study (See Methods). The clustering analysis was performed separately for each species to maximize the number of analyzed reads and hence the sensitivity and accuracy of the repeat data obtained, as described in Macas et al. (2015). To identify all satDNA families, only cluster making up at least 0.01% of the genome were analyzed due to computational constraints. Moreover, the preferential trimming of reads at 100 bp is in accordance with a previous analysis that shows that reads shorter than 80 bp tend to generate underestimated results (Sessegolo et al. 2016). The nomenclature of all new satDNAs was managed as suggested by Ruiz-Ruano et al. (2016), which include: species name abbreviation, a number in decreasing abundance order and the repeat unit size.

The overall description of all the satDNAs and repetitive DNA content of each of the 36 species are given in Figure 1, Table 1 and Table S2. Moreover, all consensus families obtained in this study are present on Supp. Material 1. Our results indicate that most of the

identified satDNA correspond to new families in *Drosophila, which* implies that a still large amount of unknown satDNA need to be discovered.

The high number of non-homologous satDNA families found in *Drosophila* reinforces the assumption that eukaryote genomes usually contain a high diversity of satDNA families. The combined proportions of these repeat clusters varied from 8.5% in *D. erecta* up to 39% in *D. montana* (Table 1). Most clusters from all species can be assigned to specific repeat types and families, except for the *D. seriema* genome that showed ~4% of bacterial contamination that was removed after manual curation.

In general, *Drosophila* satDNAs differ in nucleotide sequence, complexity, motif length, abundance and chromosome localization (Palomeque and Lorite 2008). The classification of *Drosophila* satDNAs is based according to motif length, roughly dividing it as simple or complex. Simple repeats are sequences of 5-12 bp in length, whereas complex repeats are sequences of 120-370 bp, although other satellite variants have also been found (Bonaccorsi and Lohe 1991; Gall 1974, Kuhn et al. 2008). Likewise, the 172 satDNAs showed high variation for monomer length (2–570 bp) (Table S2).The interval between 1-10 bp satDNAs contained the largest number of monomers (18), whereas complex repeats with 180-200 bp showed the second largerst number of monomers (16). This result contradicts the current hypothesis that links preferred monomer length to the length of DNA wrapped around 1 or 2 nucleosomes (~150-360 bp; Henikoff et al. 2001). Moreover, the nucleosome hypothesis suggests that 180-360 bp monomers are likely to be associated with centromeric DNA. However, simple satDNA families <50 bp (52 monomers) are more frequent than complex satDNAs (Figure 3). In addition, the extensively studied centromere of *D. melanogaster* appears to be primarily composed of short repetitive sequences (AATAT and AAGAG) with very low sequence variation (Sun et al. 2003).

Although the monomer length showed a tendency to short repeats, a major limitation is the high similarity observed among these monomers which can be clustered

together due to the cut-off parameters available (Novák et al. 2013). Previous studies have shown that several short satDNAs are present in the heterochromatic region of *D. melanogaster* and closely related species, and that these sequences share 80-90% of sequence similarity (Jagannathan et al. 2017). Thus, the underrepresentation of simple satDNAs identified in *D. melanogaster* genome can be associated with the *RepeatExplorer* limitations to clearly isolate highly similar small monomers, merging them into a single cluster. Also, we did not identify any satDNA family organized in higher-order repetition (HOR), except for the previously described *pBuM-2* α/β repeats of *D. seriema* (Kuhn et al. 2009; De Lima et al. 2017).

Secondly, our analysis showed that the GC content of satDNA in *Drosophila* is not conserved, varying from 9.1–71.8% (Table S2). Herein, we observed that most satDNA are slightly AT-rich, although several satDNA can be very GC-rich. However, it has been postulated that *Drosophila* genomes are, in general, AT-rich (*Drosophila* 12 Genomes Consortium 2007), suggesting that satDNAs may share the overall genomic configuration of non-repetitive sequences. Evenly, Melters et al. (2012) observed a similar pattern of AT/GC composition when analyzed the most abundant tandem repeats of 282 species.

Furthermore, our results showed more dissimilarities than similarities to the sequences described as most abundant by Melters et al. (2012). Despite that, our data corroborate the findings of the most abundant satDNA families from *D. albomicans, D. biarmipes, D. ficusphila*, *D. pseudoobscura, D. persimilis*, *D. rhopaloa* and *D. takahashii*. Also, we confirmed that the most abundant family of *D. ananassae* is the same present in Melters et al. (2012), although the monomer length is half the size of the previous characterization (Table S2). Regardless of these results, it is important to highlight that most of the dissimilarities are linked to the different sensitivity from both methodologies. For example, Melters et al. (2012) did not analyze sequences with monomer length under 50 bp, which comprise a significant portion of most abundant satDNA families of *Drosophila* (Table S2).

In summary, the *de novo* characterization of 36 *Drosophila* satellitome using *RepeatExplorer* pipeline has confirmed the presence of previously identified sequences and broadened the knowledge of the highly divergent nature of satDNAs in *Drosophila* species.

Coexistence of related repeats supports the satDNA library hypothesis and reveals the maintenance of *1.688* satDNA for ~27 My

The identification and sequence analysis of all satDNA evidenced that numerous satDNAs are present in parallel within genomes of related species, forming the so called satDNA library (Table S2; Fry and Salser 1977). This hypothesis assumes that different satDNA families coexist within the same genome forming a collection of repetitive sequences shared among lineages. Moreover, the abundance of particular families changes stochastically through both expansion and shrinkage (Mestrovic et al. 1998). The most remarkable case in insects is observed in the coleopteran genus *Palorus* in which four satDNA families remain completely frozen for long evolutionary periods of up to 60 My (Mravinac et al. 2005).

To confirm this hypothesis, we selected the four groups of species that were better represented in literature. Moreover, we removed from the analysis the simple satDNAs, which despite the sequence similarities, do not share significant certainty of homology. Previously, our group identified the maintenance of *pBuM* satDNA family on *D. mojavensis* and *D. buzzatii* cluster species, which shared a common ancestor about ~12 My (De Lima et al. 2017). We confirmed this data and identified that *pBuM* family is also present in the *D. arizonae* genome (Figure S1, Table S2). Cactophilic *Drosophila* seem to be the group of species with the lowest amount of satDNA (De Lima et al. 2017). However, a pool of ancestral satDNAs is maintained throughout the phylogeny of this group (Figure 4a). In this regard, a similar scenario of amplification of different satDNA variants was observed in *D. buzzatii* species cluster (Kuhn et al. 2008). In addition, we also found that despite both subspecies of *D. mojavensis* share the same satDNA families, *D. mojavensis baja* showed a

lower amount of satDNA sequences when compared to *D. mojavensis wrigleyi* (Figure 4a). The same is observed for the *D. pseudoobscura* and *D. persimilis* genomes which share the same four satDNA families, but the abundance of each satDNA family varied significantly between both species, evidencing the recurrent turnover of sequences even in close related species (Figure 4b). Accordingly, the phylogenetic close species *D. tristis*, *D. ambigua,* and *D. obscura* showed a significant variation the of satDNA family *ATOC* (Bachmann and Sperlich 1993)

We confirmed that the satDNA family *pvB370* is apparently conserved for a period of about 20 My in the *D. virilis* group species but with different abundances on each species (Heikkinen et al. 1995; Biessman et al. 2000). We also confirmed the existence of a 172 bp satDNA family previously described by Abdurashitov et al. (2013). Additionally, we described one new satDNA family with 393 bp shared among the four species of the *D. virilis* group analyzed in this study (Table S2). As expected from the satDNA library hypothesis, all four satDNA families presented different genomic proportions in all four species which diverged ~10 Mya, evidencing that non-homologous satDNA families can be preserved in the long term, yet independently amplified on each species (Figure 4c). Similarly, the recently discovered satDNA *Tetris-220* is only present in *D. virilis* and *D. americana* genomes (Dias et al. 2014), but absent on *D. novamexicana* (Figure 4b). The low abundance (or absence) of *Tetris-220* in *D. novamexicana*, closely related to *D. americana*, may be related to the overall shrinkage of satDNA in this species, the lowest among the *D. virilis* group (Figure 2).

However, the most striking coexistence of satDNA is observed on species from *D. melanogaster* group. Herein, we describe the conservation of the *1.688* satDNA family with a significant abundance on the genomes of *D. melanogaster* subgroup and *on D. takahashii, D. eugracilis, D. elegans, and D. biarmipes* (Figure 3d). These species shared a common ancestor ~27 Mya, indicating and the broader phylogenetic maintenance of a satDNA family in *Drosophila* species so far. It is also noteworthy that *1.688-like* sequences show a greater abundance on *D. elegans* when compared to the *D. melanogaster* subgroup (Figure 4d).

Furthermore, to confirm this result we ran a phylogenetic analysis of all *1.688* consensus sequences generated on the 13 species (Figure 5). Our results indicate that, despite its presence in all 13 genomes, the *1.688* family shows a significant variation, on either monomer length or nucleotide divergence (Table S3). This lack of similarity is also found on the phylogenetic reconstruction tree (Figure S3), evidencing that *1.688* sequences have evolved in disagreement with the phylogeny relationship of the species.

The presence of *1.688* arrays on *D.melanogaster-D.biarmipes* suggests that this satDNA family was present in these species common ancestor genome ~27 Mya (Russo et al. 2013). Previous analyses have shown that *1.688* heterochromatic copies of the *melanogaster* subgroup tend to evolve on a clear pattern of concerted evolution (Kuhn et al. 2012; Strachan and Dover 1985). Therefore, the high nucleotide divergence observed is probably the output of homogenization and fixation of mutations caused by molecular drive mechanisms (Dover 1986). Although we did not observe any satDNA family with similar characteristics in *D. kikkawai*, this may be either a consequence of stochastic processes or concerted evolution may have led to a significant nucleotide divergence that may hinder the identification in this species.

However, one of the most remarkable characteristic of *1.688* family is its presence throughout euchromatic arrays of chromosomes X, 2, and 3 (Losada and Villasante 1996; Kuhn et al. 2012; Khost et al. 2017; De Lima et al. in preparation). Moreover, recent analyses have evidenced that *1.688* arrays present on X chromosome euchromatin play an important role in dosage compensation of *D. melanogaster* (Menon et al. 2014). Additionally, *1.688* transcripts are also associated with Cenp-A and Cenp-C proteins being required to the kinetochore formation in this species (Rosic et al. 2014). Despite the necessity of further analysis of chromosomal mapping and transcriptional profile of these sequences to confirm these results on the other species that share this satDNA family, the presence of *1.688* arrays on a broad range of species may indicate possible functional constraints on satDNA

sequences that may be related to specific protein-binding sites (Kuhn 2015). This indication is additionally supported not only by the evolutionary conservation but also by the presence of significantly conserved and variable regions of the monomers observed throughout all 13 species (Figure 4).

It is important to highlight that the two satDNA families with broader coexistence pattern throughout *Drosophila* phylogeny, *pvB370* and *1.688*, were previously described on both domains of chromatin (Kuhn et al 2012; Biessman et al. 2000). Therefore, the maintenance of both satDNA families may also be associated with evolutionary constraints regarding to euchromatic regions of the genome (Kuhn et al. 2012).

TEs structural sequences display a minor influence on satDNAs origin in *Drosophila*

Despite conceptual differences, there are many reports showing that TEs can contribute to the origin and/or amplification/homogenization of satDNA. A clear relationship between TEs and the origin or amplification/homogenization of satDNA has been suggested in several insect species, especially in *Drosophila* (Palomeque and Lorite 2008; McGurk and Barbash 2017). Our findings support the previous description of three tandemly repeated families present on *D. virilis* group: *pvB370* satellites, which originated from *pDv* elements; *Tetris-220*, originated from a *Foldback*-like element; and 150 bp repeats, which are associated with *DINE-TR1* CTRs (Heikkinen et al. 1995; Dias et al. 2014; Dias et al. 2015, respectively). Herein, a comparative analysis showed that *pvB370* is widely present on *D. virilis* group, as described by Biesmann et al. (2000), whereas *Tetris-220* and *DINE-TR1* satDNA-like sequences are constrained to *D. virilis* and *D. americana* genomes. All these satDNAs showed sequence and abundance variation (Figure 3b).

We also identified two new satDNA families on *D. biarmipes* genome and one family on *D. pseudoobscura* group species (*D. pseudoobscura* and *D. persimilis*) that shared sequence similarity to previously described TEs. The first one, *DbiaSat2-263*, was previously deposited on RepBase as *Copia-2_DTa-I_1p*, whereas *DbiaSat5-215* was reported as part of

the *Gypsy-11_DEu-I*. Aiming to describe the sequence origin, we run a scrutinized analysis and evidenced that both satDNA families, *DbiaSat2-263* and *Dbia5-215*, do not show similarity with any structural fragments of *Copia* and *Gypsy* TEs families, respectively. Likewise, we found that the *Gypsy18-LTR_Dpse* sequence deposited on RepBase database is, in fact, a 21 bp satDNA family that forms large arrays and isshared among *D. pseudoobscura* and *D. persimilis*. Moreover, we have identified that the *Gypsy18-I_Dpse* have a 228 bp LTR sequence inside the previously described element, indicating an incorrect characterization of this specific TE sequence.

Our group has previously identified that the *CDSTR130* satDNA family present on *D. mojavensis* genome was previously described on RepBase as a Long Terminal Repeats (LTR) of the *BEL3_DM-I* element. However, a manual curation has shown that this 130 bp tandemly repeated family is not part of the LTR but only flank the element, thus indicating an incorrect annotation of the *BEL3_DM_LTR* sequence (De Lima et al. 2017).

To confirm our characterization, we ran BLAST searches using satDNA consensus sequences (*DbiaSat2-263*; *Dbia5-215*; *DperSat3-21* and *DpseSat2-21*) against the *D. biarmipes* and *D. pseudoobscura* group species genomes to obtain tandemly repeated arrays of each family. We analyzed the BLAST results that presented the highest score number, simultaneously to the preference of single contigs composed by each satDNA family. We identified a single array with ~20 monomers of the *DbiaSat2-263* family present in tandem on Contig5744 (AFFD02005737.1). We also observed that *DbiaSat5-215* shows several contigs composed exclusively of the 215 bp monomers (e.g. AFFD02001879.1; AFFD02002683.1). The same pattern is also observed for *DperSat3-21* and *DpseSat2-21* satDNA on *D. pseudoobscura* and *D. persimilis* assembled genomes (AAIZ01026166.1; AAIZ01022678.1; AADE01010412.1).

Thus, our results suggest that both satDNA families identified in *D. biarmipes* and the one found in *D. pseudoobscura* group species do not share homology with TEs fragments,

being more likely to be an incorrect annotation of TE elements. Altogether, satDNA originated from TEs are particularly common on *D. virilis* group species, but are mostly absent on the other analyzed species.

Drosophila satDNA content and genome size evolution hypotheses

Given the ~2.5-fold variation in genome size observed in the analyzed *Drosophila* species, we expected to find evidence of repetitive elements substantially contributing to nuclear DNA content (Gregory and Johnston 2008). The satDNA abundance and the overall repetitive DNA content of each species of 36 species are given in Table 1. Genome size of *Drosophila* species used in this analysis range from 139.9 Mb to 325.4 Mb with an average of 196.2 Mb and a median of 193. The variation in satDNA content and its significance for changes in genome size is consistent with previous ideas that genomes have expanded/contracted mainly by addition/deletion of repeated sequences (for review see Gregory 2005).

Our estimates show a much lower genome proportion of satDNA sequences when compared to previous analyses in *Drosophila* (Bosco et al. 2007; Craddock et al. 2016). These analyses proposed that heterochromatic sequences do not replicate during polyploid follicle cells development and used the percentage of under-replication to estimate the heterochromatic satDNA content in each species (Bosco et al. 2007). However, this approach does not differentiate the satDNA sequences from the heterochromatin itself, and do not embrace interspersed organization of satDNA repeats arrays interrupted by several TEs (Khost et al 2017). Thus, this artifact may overestimate the actual genomic contribution of satDNA sequences, suggesting that the significant difference observed between the two methodologies may be associated with several TE insertions or other non-satDNA sequences within constitutive heterochromatin, as observed in *D. melanogaster* (Khost et al. 2017; Sun et al. 2005).

The satDNA estimates observed herein reinforce previous assumptions that there is a trend in which larger genomes have more satDNA content in general (Bosco et al. 2007). For instance, *D. virilis* group species, with the largest genome mean (287 Mb) also have the highest satDNA contents (19.4-27.4%; Table 1). Consonantly, *D. repleta* group species have the smallest genome mean (155.5 Mb) also show the lowest amount of satDNA (Figure 2/ Table1). Despite the large difference in genome size and satDNA amount, both species group are classified on *D. virilis-repleta* radiation species and shared a common ancestor about 27 My (Russo et al. 2013).

Previous studies have shown that *Drosophila* genome size evolution seems to follow a significant phylogenetic signal (Craddock et al. 2016; Hjelmen and Johnston 2017). Likewise, our estimates in 36 *Drosophila* support this claim and show that the panorama of satDNA distribution is essentially shared among closely related species (Figure 2). Despite that, it is noteworthy that we observed few examples of large variations of satDNA content among short evolutionary ranges. For instance, the closely related species *D. orena, D. erecta, D. yakuba D. teissieri,* and *D. santomea* share a common ancestor ~8.2 My, yet *D. orena* shows 6-10 fold more satDNA sequences than the other four species. As consequence, *D. orena* has also the largest genome on *D. melanogaster* subgroup (280.7 Mb), almost 100 Mb larger (Figure 2). Alternatively, *D. mexicana* has the lowest amount of satDNA sequences of the subgroup *virilis* (6.82%), indicating shrinkage of these sequences when compared to the closely related species.

Regardless both previous cases of *D. orena* and *D. novamexicana*, the general pattern of satDNA distribution seems to show a phylogenetic trend (see below), suggesting that the satDNA abundance seems to evolve in a gradual manner in *Drosophila*. Contradictory, the remarkable difference of satDNA abundance observed between cactophilic flies from *D. repleta* group (lowest) and *D. virilis* group species (highest) (Figure 2; Table 1) call into question the mechanisms acting on these genomic traits. Both groups are

members of *virilis-repleta* radiation and shared a common ancestor about ~27 My, yet they exhibit a great divergence on ecological behavior and in the utilization substrates (Markow and O'Grady 2005). Members of the *D. virilis* group occupy a wide range of substrate types, are found on all continents and display a greater diversity of resource specializations. Oppositely, *D. repleta* species are mostly restricted to desertic areas and have a high association level with cactus (Oliveira et al. 2012). Therefore, it is tempting to suggest that this major variation on genome size and satDNA abundance between sister clades, such as *D. virilis* and *D. repleta*, may be associated with environmental adaptations, development times or oviposition and developing in ephemeral hosts (Markow and O'Grady 2008).

The current dichotomy between adaptive and junk DNA theories focus objectively on the question of whether repetitive DNA benefits the organism and if genome size variation resulted from repetitive DNA expansion and influenced by natural selection (Petrov 2001; Graur et al. 2015). The adaptive hypothesis proposes that genome size evolution should track adaptive needs of the efficacy of natural selection on different organisms (Gregory 2005; Gregory and Johnston 2008). The pattern observed on subgenus *Drosophila*, especially on the *virilis-repleta* radiation, indicates the presence of a significant phylogenetic signal influencing the genome size evolution throughout *Drosophila*. This statement suggests that genome size variation does not fit well with the adaptive hypothesis (Hjelmen and Johnston 2017). Moreover, the adaptive explanation itself does not seem to hold for *Drosophila virilis-repleta* species, since *D. virilis* and *D. mojavensis* do not have an embryonic developmental rate that correlates significantly with their great differences in genome size and body size (Markow et al. 2009). Moreover, species with large genomes, such as *D. virilis* and *D. funebris* have a slow developmental rate and large body size, yet, *D. willistoni*, with a genome size nearly identical to *D. funebris*, is a rapidly developing small species (Gregory and Johnston 2008).

Oppositely, it is argued that selection is ineffective at lower effective population sizes, and therefore potentially maladaptive changes in genome size may accumulate and persist

in the population (Lynch and Conery 2003) Likewise, the rate of genome size change on *Sophophora* species is faster on the early stages of phylogeny with a decrease in that rate as the groups differentiation went on through (Hjelmen and Johnston 2017). Therefore, during the process of the *D. virilis-repleta radiation,* it is plausible that the small effective population sizes on initial speciation process were critical to the differential fixation of satDNA abundance. Accordingly, it is suggested that population bottlenecks and founder effect speciation in island fauna influenced the increase in the genome size and satDNA sequences in the Hawaiian *Drosophila* when compared to non-Hawaiian *Drosophila* species (Craddock et al. 2016).

Altogether, we observed that satDNA content is highly variable among *Drosophila* (Table 1). However, this variation seems to have a phylogenetic signal that is likely to be related to repetitive DNA itself (detailed in Dodsworth et al. 2014; Macas et al. 2015) and that correlates with genome size, but currently is not possibleto determine if these patterns are the result of environmental pressures leading to rapid genome size change in *Drosophila* species.

SatDNA and TEs have a distinct influence on the genome size evolution of *Sophophora* and *Drosophila* subgenera

Prompted by the results described above, we ran statistical analyses to investigate how the global variability of satDNA families and repetitive DNAs influence the genome size evolution in *Drosophila* (Figure 6). An overall correlation of genome size versus satDNA content showed a significant positive relationship among the 36 species (Spearman's correlation coefficient r=0.654, p=0.000036) (Table 2). Furthermore, the overall repetitive content has a significant correlation with the genome size variation observed in *Drosophila* (r=0.658, p=0.000032). Interestingly, the correlation between satDNAs and overall repetitive sequences is smaller than the expected by the previous analysis and show only a moderate

value (*r*=0.513, *p*=0.001158), suggesting that these variables may dissimilarly influence the genome evolution of *Drosophila* species.

This incongruence raised the question if the genome size variation of each subgenera is influenced equally by satDNA and overall repetitive content. To correct for phylogenetic groups, we run statistical analyses using each subgenera data independently. The positive relationship between genome size and satDNA is strongly significant on the *Drosophila* subgenus even after correcting for phylogeny (*r* =0.947, *p*=0.000030). The same was observed for satDNA and repetitive content (*r* =0.867, *p*=0.00026) and the correlation between repetitive content and genome size variation on *Drosophila* subgenus (r=0.658, *p*= 0.038404). Oppositely, the satDNA abundance in *Sophophora* species seems to have a small influence on the genome size variation observed in the subgenus (*r*=0.380, *p*=0.073328). Likewise, the correlation between the overall repetitive sequences and satDNA abundance is very low (*r*=0.205, *p*=0.324673), suggesting that other repetitive families are the major components of *Sophophora* genomes. Indeed, the repetitive content and genome variation in *Sophophora* have the highest *r* correlation observed for both variables (*r*=0.712, *p*=0.000137) (Table 2).

The genome size variation of *Drosophila* is assumed to be associated with the amplification/contraction of the heterochromatic blocks (Gregory and Johnston 2008; Craddock et al. 2016), mostly composed by satDNAs and TEs arrays. Properly, our results show that both TEs and satDNA can influence genome size in *Drosophila*. Ergo, among the species with the smallest genomes (150 Mb or smaller), such as *D. arizonae, D. subobscura* and *D. busckii*, none have more than 2% satDNA or more than 16% of repetitive DNA (Table 1).

Nevertheless, the estimate of the overall repetitive DNA and satDNA content do not seem to influence homogeneously the genome size evolution of both subgenera. As described in Table 2, genome size variation on *Drosophila* subgenus is mostly related to the expansion/shrinkage of the satDNA sequences, but this association is not found in

Sophophora genomes. Moreover, the sum of all repetitive DNA identified in *Sophophora* does not hold correlation with the satDNA abundance described (Table 2) as notably illustrated in *D. affinis* and *D. takahashii* genomes (Figure 1 and 2). This particular result suggests that non-satDNA repetitive DNAs are probably responsible for the genome size variation observed on *Sophophora*. Indeed, it has been recently shown that TEs accumulation emerges as a major factor of genome size variation in *Sophophora* species, especially on *D. melanogaster* subgroup (Sessegolo et al. 2016). Similarly, the substantial expansion of the majorly heterochromatic *D. ananassae* Muller F element (dot chromosome) is directly associated with a burst of LTR and LINE retrotransposons that comprise 78.6% of *D. ananassae* F element (Leung et al. 2017). Additionally, the distribution of TEs was associated toward the accumulation of larger genomes in different populations of *D. melanogaster* (Huang et al. 2014).

Despite *D. orena* genome size being linked to the amplification of one satDNA family and that *D. suboobscura* small genome has a low amount of satDNA sequences (Figure 2), it is reasonable to infer that TEs do have a greater influence on *Sophophora* genomes than satDNAs. One straightforward explanation is that TEs frequency in *Sophophora* has increased substantially since the divergence of this group from the *Drosophila* subgenus. However, this hypothesis does not explain the lower influence of satDNA and TEs on the genome size evolution of *Sophophora* and *Drosophila* subgenera, respectively. Moreover, it is important to highlight that different molecular mechanisms, like the size and frequency of small spontaneous nucleotide insertions and deletions (indels) or the length of introns, can also be important in the long-term evolution of genome size in *Drosophila*, regardless the repetitive DNA abundance (Kelly et al. 2015; Gregory 2003; Petrov and Hartl 1998; Moriyama et al. 1998).

Altogether, our results agree with the hypothesis that genome size evolution in *Drosophila* can be differently influenced by satDNA and TEs. Moreover, it also suggests that

Sophophora and *Drosophila* subgenera genome size evolution occurs under different scenarios and that broader analysis of repetitive sequences is necessary to fully explain the strength, causes or effects of satDNA sequences on the modulation of genome size in *Drosophila*.

Concluding Remarks

Herein we performed the largest satellitome analysis characterization of *Drosophila* species. In summary, the *de novo* characterization of 36 *Drosophila* satellitome using *RepeatExplorer* pipeline has confirmed the presence of previously identified sequences and described 133 new satDNA families, broadening the scarce knowledge of the highly divergent satDNA sequences found throughout the *Drosophila* phylogeny. We described that satDNA sequences have a highly variable content on *Drosophila*, yet this variation seems to follow phylogenetic relationships, and the closely related species tend to show a similar amount of satDNA sequences. Despite that, we also observed multiple independent turnovers events that have changed the abundance of satDNA families over evolutionarily short timescales. Notably, we have described that the *1.688* satDNA family is much more ancient than it was predicted, as it is shared among species that diverged for at least 27 My.

Moreover, the description of the satDNA content of each species concomitantly to the use of previously described genome size data helped to look more closely the direct influence of satDNA on genome size evolution on *Drosophila* species. In contrast to previous studies, we provide evidence that genome size evolution of both *Drosophila* subgenera is differently influenced by satDNA and TEs. We observed that *Drosophila* subgenus genomes are strongly shaped by satDNA expansion/contractions, whereas *Sophophora* species seem to hold a greater correlation with TEs. Finally, the identification of 133 new satDNA sequences of 36 species of *Drosophila* engenders significantly the characterization of repetitive sequences, improving the overall use of these sequences on genome assemblies.

List of Figures:

Figure 2. Genome size evolution and repeat composition of 36 *Drosophila* species and one subspecies of *D. mojavensis*. Repetitive DNA (red) and Satellite DNA (blue) proportions in Mb of *Drosophila* species calculated according to each species genome size (green).

Figure 3. A. Monomer length of the 172 satellite DNA sequences described in the 36 species of *Drosophila* separated on 10 bp intervals. **B.** GC content variation of the monomers identified on 10% intervals. **C.** The overall distribution of satDNA families genomic proportion analyzed independently.

Figure 4. A-C. Variation in satDNA library profile among close related species from **(A)** *D. repleta* group; **(B)** *D. virilis* group; **(C)** *D. pseudoobscura* group. **D**. Variation in *1.688* satDNA abundance on 13 species of *D. melanogaster* group.

Figure 5. Representative sequence alignment of *1.688* consensuses of 13 *Drosophila* species described in this study indicating the presence of conserved regions. Dark blue indicates regions with higher sequence conservation while white regions indicate no conservation among the sequences.

Figure 6. Correlation of repeats with genome size. Graphs show the Spearman's rank correlation between Repetitive Content and Genome Size; Repetitive Content and SatDNA Content; and SatDNA Content and Genome Size in **(a)** *Drosophila* genus, **(b)** Sophophora subgenus and **(c)** *Drosophila* subgenus.

List of Tables:

Table 1: Repetitive content estimation and Satellite DNA contribution of 36 *Drosophila* species.

Table 2: Spearman's correlation coefficient among Repetitive DNA content, satDNA content and genome size variation in *Drosophila* genus, and both subgenus Sophophora and *Drosophila* independently.

Supplementary Figures

Figure S1. A-C. Sequence alignment of three complex satDNA families shared among species from *D. virilis* group. **D.** Sequence alignment of *pBuM* satDNA family consensus shared among *D. arizonae, D. mojavensis* and *D. buzzatii* genomes.

Figure S2. Sequence alignment of three complex satDNA families shared between *D. pseudoobscura* and *D. persimlis* genomes.

Figure S3. Maximum Likelihood phylogenetic relationship among 13 consensus sequences of *1.688* satellite DNA family found on *melanogaster* group species. The tree was constructed with T92+G evolutionary model of substitution and 1,000 replicates of bootstrap.

Figure S4. Boxplots of significance values from (a)Repetitive DNA and (b) Satellite DNA contents identified on both subgenus of *Drosophila*

Supplementary Tables:

Table S1. Estimation of genome size used in this study and a total number of 100bp Illumuina reads used for each species and the respective genome coverage.

Table S2. Description of all satDNA families identified by *RepeatExplorer* pipeline in the present study from 36 species of *Drosophila*.

Table S3. Interespecific genetic distances (*p*-distance) among *1.688* satellite DNA consensus sequences of 13 species of *D. melanogaster* group.

Supplementary Material

Supplementary Material 1. Nucleotide sequence consensus from all satDNA families identified in the present study. Nomenclature of all new satDNA families were done according to Ruiz-Ruano et al. (2016).

References

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

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

Bachmann, L., and Sperlich, D. (1993). Gradual evolution of a specific satellite DNA family in *Drosophila* ambigua, D. tristis, and D. obscura. *Molecular biology and evolution*, *10*(3), 647-659.

Bachtrog, D. (2003). Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. *Nature genetics*, *34*(2), 215.

Bachtrog, D. (2005). Sex chromosome evolution: molecular aspects of Y-chromosome degeneration in *Drosophila*. *Genome research*, *15*(10), 1393-1401.

Bachtrog, D. (2013). Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nature Reviews. Genetics*, *14*(2), 113.

Barnes, S. R., Webb, D. A., and Dover, G. (1978). The distribution of satellite and main-band DNA components in the *melanogaster* species subgroup of *Drosophila*. *Chromosoma*, *67*(4), 341-363.

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

Biessmann, H., Zurovcova, M., Yao, J. G., Lozovskaya, E., and Walter, M. F. (2000). A telomeric satellite in *Drosophila virilis* and its sibling species. *Chromosoma*, *109*(6), 372-380.),

Bonaccorsi, S., and Lohe, A. (1991). Fine mapping of satellite DNA sequences along the Y chromosome of *Drosophila melanogaster*: relationships between satellite sequences and fertility factors. *Genetics*, *129*(1), 177-189.

Bosco, G., Campbell, P., Leiva-Neto, J. T., and 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.

Brutlag, D., Appels, R., Dennis, E. S., and Peacock, W. J. (1977). Highly repeated DNA in *Drosophila melanogaster*. *Journal of molecular biology*, *112*(1), 31-47.

Cattani, M. V., and Presgraves, D. C. (2012). Incompatibility between X chromosome factor and pericentric heterochromatic region causes lethality in hybrids between *Drosophila melanogaster* and its sibling species. *Genetics*, *191*(2), 549-559.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, *371*(6494), 215-220.

Craddock, E. M., Gall, J. G., and Jonas, M. (2016). Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. *Genetica*, *144*(1), 107-124.

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

de Lima, L. G., Svartman, M., and Kuhn, G. C. (2017). Dissecting the Satellite DNA Landscape in Three Cactophilic *Drosophila* Sequenced Genomes. *G3: Genes, Genomes, Genetics*, g3-117.

Dias, G. B., Svartman, M., Delprat, A., Ruiz, A., and 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.

Dias, G. B., Heringer, P., Svartman, M., and 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.

Dodsworth, S., Chase, M. W., Kelly, L. J., Leitch, I. J., Macas, J., Novák, P., ... and Leitch, A. R. (2014). Genomic repeat abundances contain phylogenetic signal. *Systematic biology*, *64*(1), 112-126.

Dover, G. A., and Tautz, D. (1986). Conservation and divergence in multigene families: alternatives to selection and drift. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *312*(1154), 275-289.

Drosophila 12 genomes Consortium 2007; Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, *450*(7167), 203.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.

Ferree, P. M., and Barbash, D. A. (2009). Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS biology*, *7*(10), e1000234.

Ferree, P. M., and Prasad, S. (2012). How can satellite DNA divergence cause reproductive isolation? Let us count the chromosomal ways. *Genetics research international*, *2012*.

Fry, K., and Salser, W. (1977). Nucleotide sequences of HS-α satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. *Cell*, *12*(4), 1069-1084.

Gall, J. G., Cohen, E. H., and Polan, M. L. (1971). Repetitive DNA sequences in *Drosophila*. *Chromosoma*, *33*(3), 319-344.

Gall, J. G., and Atherton, D. D. (1974). Satellite DNA sequences in *Drosophila virilis*. *Journal of molecular biology*, *85*(4), 633-664.

Gallach, M. (2014). Recurrent turnover of chromosome-specific satellites in *Drosophila*. *Genome biology and evolution*, *6*(6), 1279-1286.

Garavís, M., Méndez-Lago, M., Gabelica, V., Whitehead, S. L., González, C., and Villasante, A. (2015). The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. *Scientific reports*, *5*.

Graur, D., Zheng, Y., and Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome biology and evolution*, *7*(3), 642-645.

Gregory, T. R. (2003). Is small indel bias a determinant of genome size?. *TRENDS in Genetics*, *19*(9), 485-488.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics*, *6*(9), 699.

Gregory, T. R., and Johnston, J. S. (2008). Genome size diversity in the family Drosophilidae. *Heredity*, *101*(3), 228.Guillén et al. 2015

Heikkinen, E., Launonen, V., Müller, E., and Bachmann, L. (1995). The *pvB370* BamHI satellite DNA family of the *Drosophila virilis* group and its evolutionary relation to mobile dispersed genetic pDv elements. *Journal of molecular evolution*, *41*(5), 604-614.

Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, *293*(5532), 1098-1102.

Hjelmen, C. E., and Johnston, J. S. (2017). The mode and tempo of genome size evolution in the subgenus Sophophora. *PloS one*, *12*(3), e0173505.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome research*, *9*(9), 868-877.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., ... and Magwire, M. M. (2014). Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome research*, *24*(7), 1193-1208.

Jagannathan, M., Warsinger-Pepe, N., Watase, G. J., and Yamashita, Y. M. (2017). Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3: Genes, Genomes, Genetics*, *7*(2), 693-704.

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

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P. and Nichols, R. A. (2015). Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytologist, 208(2), 596-607.

Khost, D. E., Eickbush, D. G., and Larracuente, A. M. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome research*, *27*(5), 709-721.

Kuhn, G. C. S., Bollgönn, S., Sperlich, D., and Bachmann, L. (1999). Characterization of a species-specific satellite DNA of *Drosophila* buzzatii. *Journal of Zoological Systematics and Evolutionary Research*, *37*(2), 109-112.

Kuhn, G. C., Sene, F. M., Moreira-Filho, O., Schwarzacher, T., and Heslop-Harrison, J. S. (2008). 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*, *16*(2), 307-324.

Kuhn, G. C., Küttler, H., Moreira-Filho, O., and Heslop-Harrison, J. S. (2011). The *1.688* repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Molecular biology and evolution*, *29*(1), 7-11.

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

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution*, *33*(7), 1870-1874.

Laird, C. D., and McCarthy, B. J. (1968). Magnitude of interspecific nucleotide sequence variability in *Drosophila*. *Genetics*, *60*(2), 303.

Larracuente, A. M. (2014). The organization and evolution of the Responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. *BMC evolutionary biology*, *14*(1), 233.

Leung, W., Shaffer, C. D., Chen, E. J., Quisenberry, T. J., Ko, K., Braverman, J. M., ... and Robic, S. (2017). Retrotransposons are the major contributors to the expansion of the *Drosophila* ananassae Muller F element. *G3: Genes, Genomes, Genetics*, g3-117.

Losada, A., and Villasante, A. (1996). Autosomal location of a new subtype of *1.688* satellite DNA of *Drosophila melanogaster*. *Chromosome Research*, *4*(5), 372-383.

Lynch, M., and Conery, J. S. (2003). The origins of genome complexity. *science*, *302*(5649), 1401-1404.

Macas, J., Novak, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., ... and Leitch, I. J. (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. *PLoS One*, *10*(11), e0143424.

Macknight, R. H., 1939 The sex-determining mechanism of *Drosophila* miranda. Genetics 24: 180–201

Markow, T. A., and O'Grady, P. M. (2005). Evolutionary genetics of reproductive behavior in *Drosophila*: connecting the dots. *Annu. Rev. Genet.*, *39*, 263-291.

Markow, T. A., and O'grady, P. (2008). Reproductive ecology of *Drosophila*. *Functional Ecology*, *22*(5), 747-759.

Markow, T. A., Beall, S., and Matzkin, L. M. (2009). Egg size, embryonic development time and ovoviviparity in *Drosophila* species. *Journal of evolutionary biology*, *22*(2), 430-434.

McGurk, M. P., and 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, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., ... and Garcia, J. F. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology*, *14*(1), R10.

Menon, D. U., Coarfa, C., Xiao, W., Gunaratne, P. H., and Meller, V. H. (2014). siRNAs from an X-linked satellite repeat promote X-chromosome recognition in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, *111*(46), 16460-16465.

Mestrović, N., Plohl, M., Mravinac, B., and Ugarković, D. (1998). Evolution of satellite DNAs from the genus Palorus--experimental evidence for the" library" hypothesis. *Molecular biology and evolution*, *15*(8), 1062-1068.

Moriyama, E. N., Petrov, D. A., and Hartl, D. L. (1998). Genome size and intron size in *Drosophila*. *Molecular biology and evolution*, *15*(6), 770-773.Moriyama et al. 1998

Mravinac B, Plohl M, Mestrovic N, Ugarkovic D (2002) Sequence of PRAT satellite DNA 'frozen' in some coleopteran species. J Mol Evol 54: 774–783

Nijman, I. J., and Lenstra, J. A. (2001). Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *Journal of Molecular Evolution*, *52*(4), 361-371.

Novák, P., Neumann, P., and Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, *11*(1), 378.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., and Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792-793.

Ohno, S. (1972). So much" junk" DNA in our genome. In *Brookhaven symposia in biology* (Vol. 23, pp. 366-370).

Oliveira, D. C., Almeida, F. C., O'Grady, P. M., Armella, M. A., DeSalle, R., and Etges, W. J. (2012). Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila* repleta species group. *Molecular Phylogenetics and Evolution*, *64*(3), 533-544.

Palomeque, T., and Lorite, P. (2008). Satellite DNA in insects: a review. *Heredity*, *100*(6), 564.

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *TRENDS in Genetics*, *17*(1), 23-28.

Petrov, D. A., and Hartl, D. L. (1998). High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Molecular Biology and Evolution*, *15*(3), 293-302.

Plohl, M., Meštrović, N., and Mravinac, B. (2012). Satellite DNA evolution. In *Repetitive DNA* (Vol. 7, pp. 126-152). Karger Publishers.

Renkawitz, R. (1979). Isolation of twelve satellite DNAs from *Drosophila* hydei. *International Journal of Biological Macromolecules*, *1*(3), 133-136.

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

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., and Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific reports*, *6*.

Russo, C. A., Mello, B., Frazão, A., and Voloch, C. M. (2013). Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zoological Journal of the Linnean Society*, *169*(4), 765-775.

Schmidt, T., and Heslop-Harrison, J. S. (1998). Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plant Science*, *3*(5), 195-199.

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

Slamovits, C. H., Cook, J. A., Lessa, E. P., and Susana Rossi, M. (2001). Recurrent amplifications and deletions of satellite DNA accompanied chromosomal diversification in South American tuco-tucos (genus Ctenomys, Rodentia: Octodontidae): a phylogenetic approach. *Molecular Biology and Evolution*, *18*(9), 1708-1719.

Steinemann, M., and Steinemann, S. (1998). Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of *Drosophila* miranda a model for sex chromosome evolution. In *Mutation and Evolution* (pp. 409-420). Springer Netherlands.

Sun X, Le HD, Wahlstrom JM, Karpen GH (2003) Sequence analysis of a functional *Drosophila* centromere. Genome Res 13:182–194

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

Walsh J.B. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics. 1987;115:553–567

Waring, G. L., and Pollack, J. C. (1987). Cloning and characterization of a dispersed, multicopy, X chromosome sequence in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, *84*(9), 2843-2847.

Zhou, Q., Ellison, C. E., Kaiser, V. B., Alekseyenko, A. A., Gorchakov, A. A., and Bachtrog, D. (2013). The epigenome of evolving *Drosophila* neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS biology*, *11*(11), e1001711.

Acknowledgments

This work was supported by a grant from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (grant number APQ-01563-14) to G.C.S.K. L.G.d.L. was supported by a doctoral fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Figure 1.

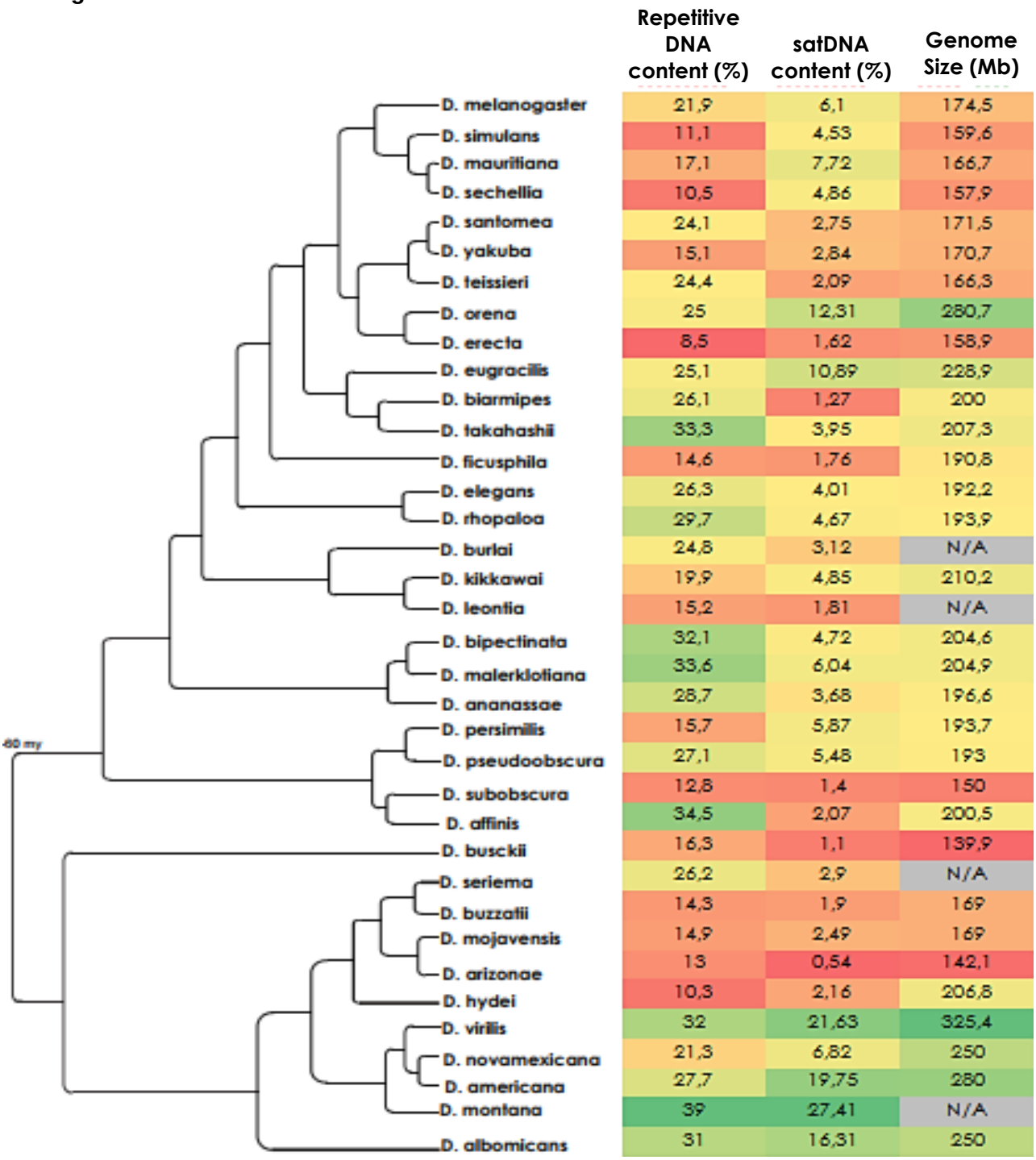

Figure 1. Overall proportion of repetitive DNA and satellite DNA amount of the 36 *Drosophila* species genomes. The intensity of green (higher) and red (lower) colors are proportional to the variation inside each column. The species are presented according to the phylogenetic tree topology as proposed by Russo et al. (2013). The genome sizes of each species are indicated.

Figure2

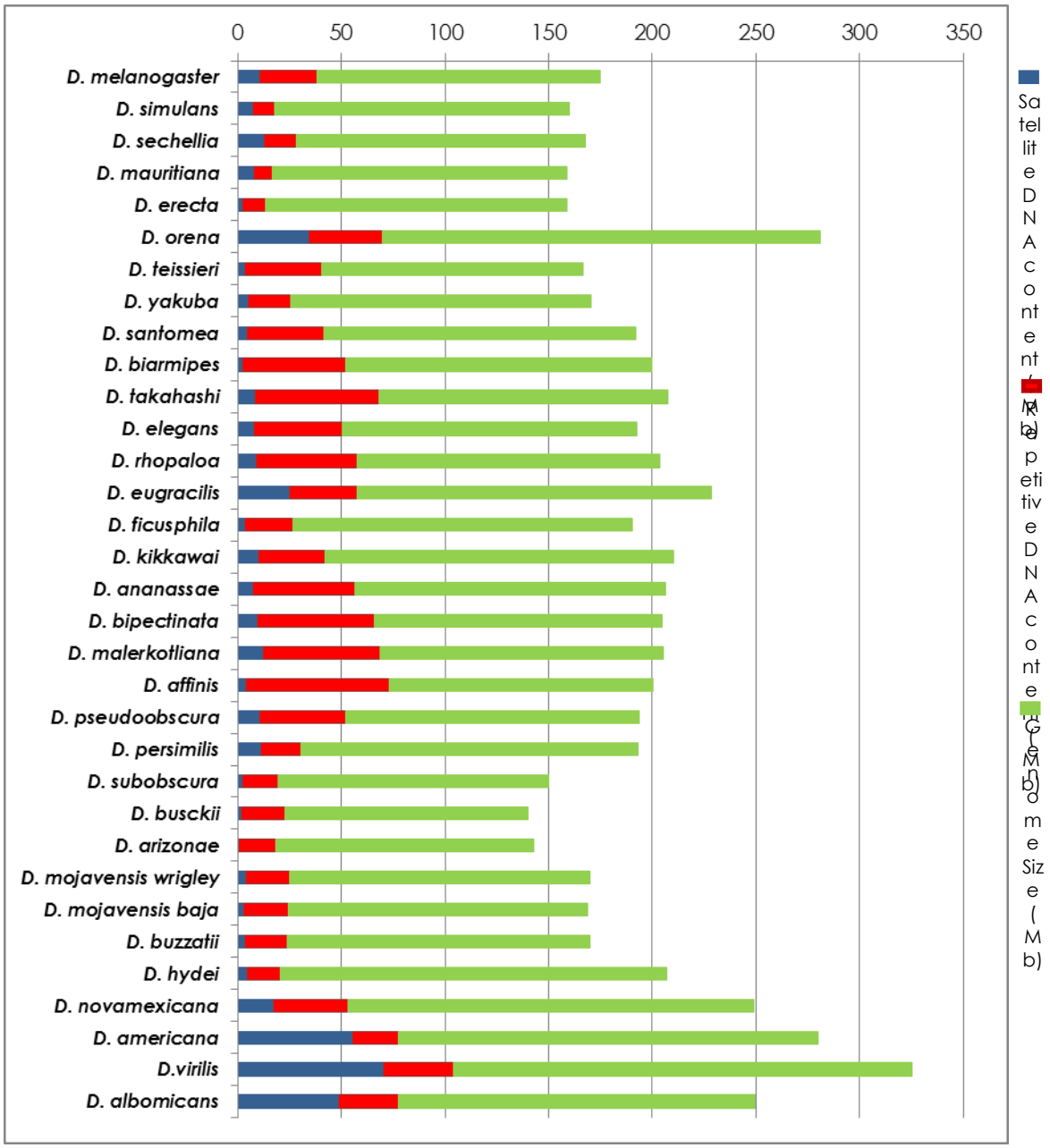

Figure 2. Genome size evolution and repeat composition of 36 *Drosophila* species and one subspecies of *D. mojavensis*. Phylogenetic tree of the *Drosophila* species investigated with their genome sizes (green) and the overall repetitive content (red) and satDNA content (blue).

Figure 3

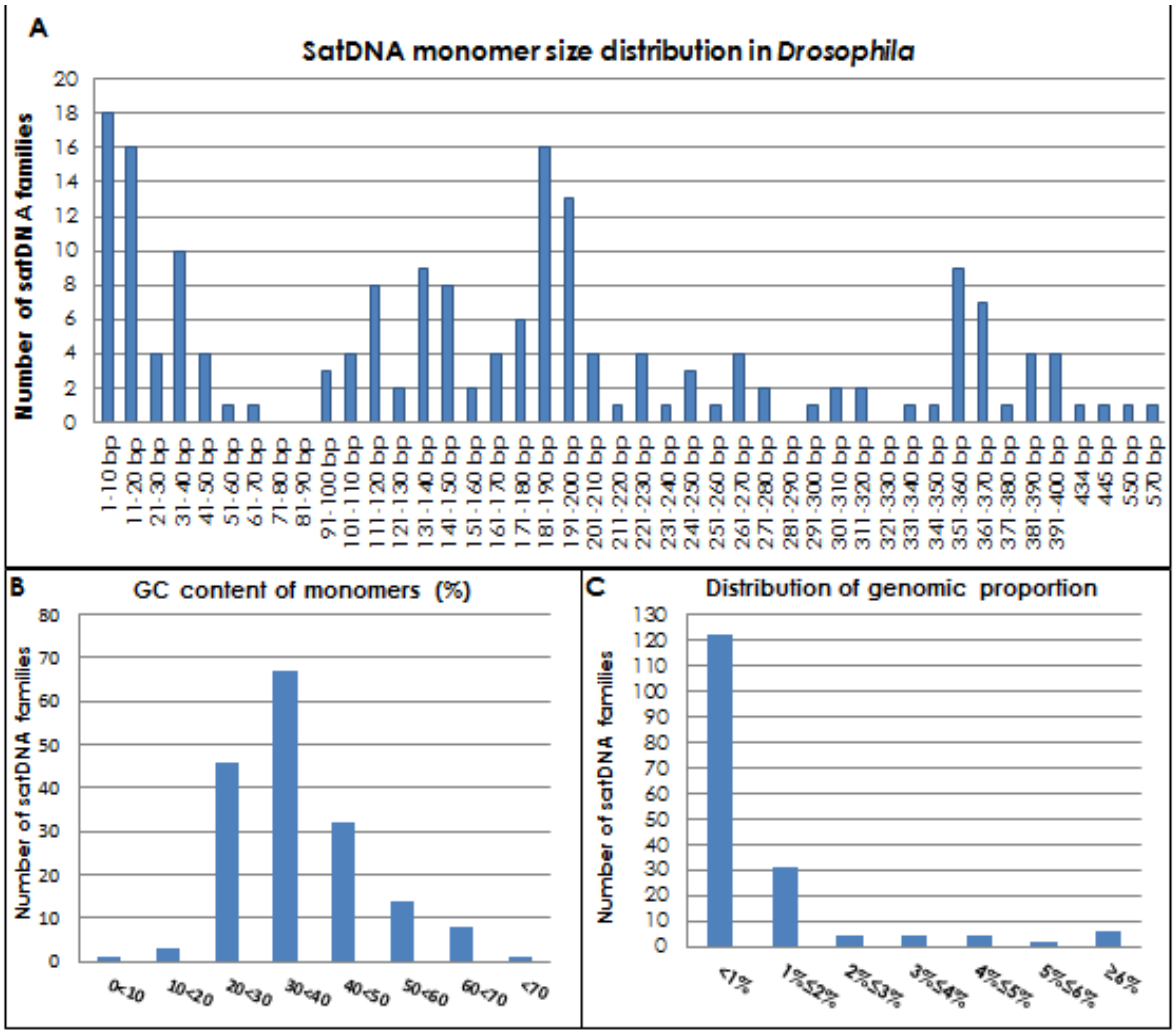

Figure 3. A. Monomer length of the 172 satellite DNA sequences described in the 36 species of *Drosophila* separated on 10 bp intervals. **B.** GC content variation of the monomers identified on 10% intervals. **C.** Overall distribution of genomic proportion of 172 satDNA families identified in 36 *Drosophila* species.

Figure 4.

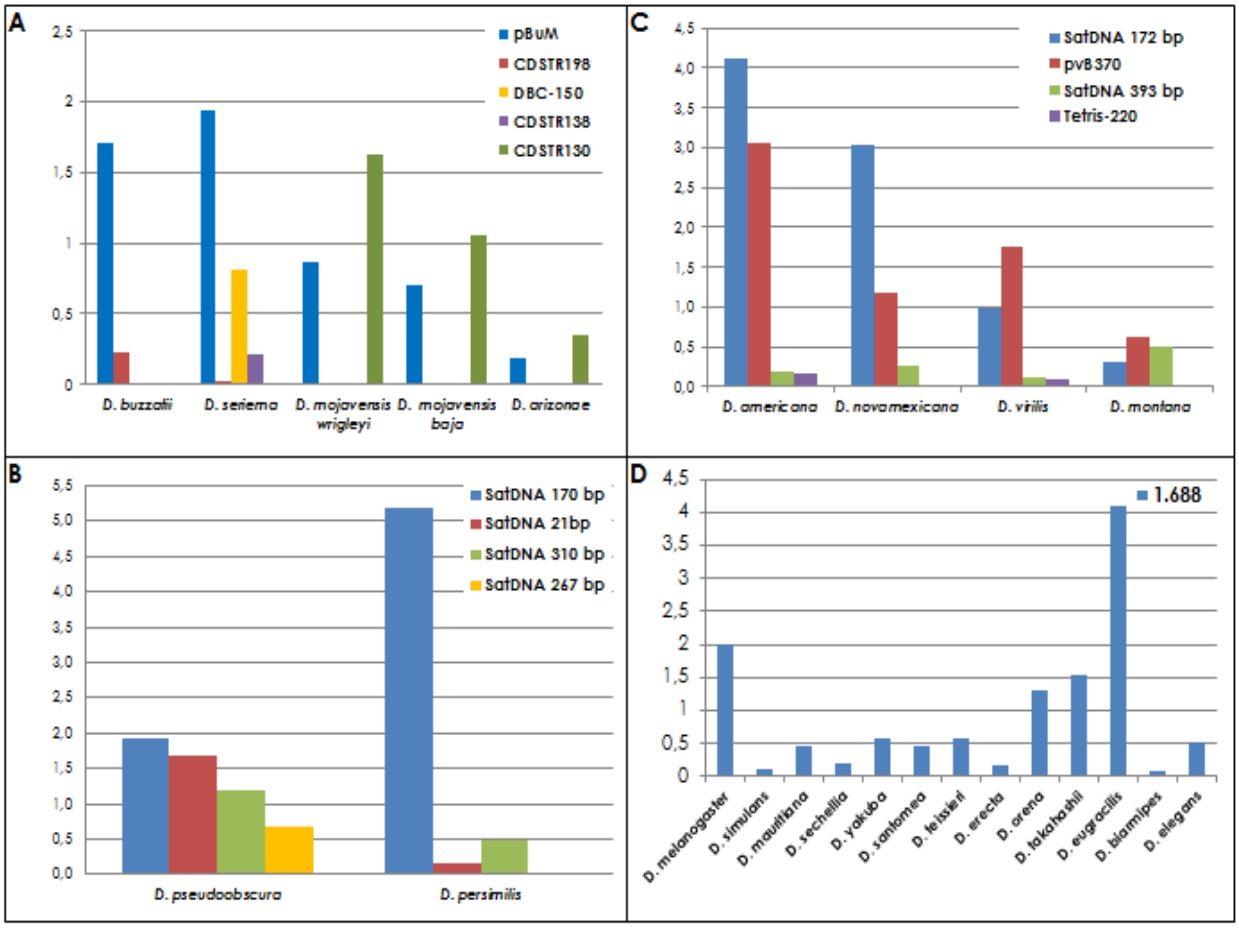

Figure 4. A-C. Variation in satDNA library profile among close related species from: **(A)** *D. repleta* group; **(B)** *D. virilis* group; **(C)** *D. pseudoobscura* group. **D**. Variation in *1.688* satDNA abundance on 13 species of *D. melanogaster* group.

Figure 5.

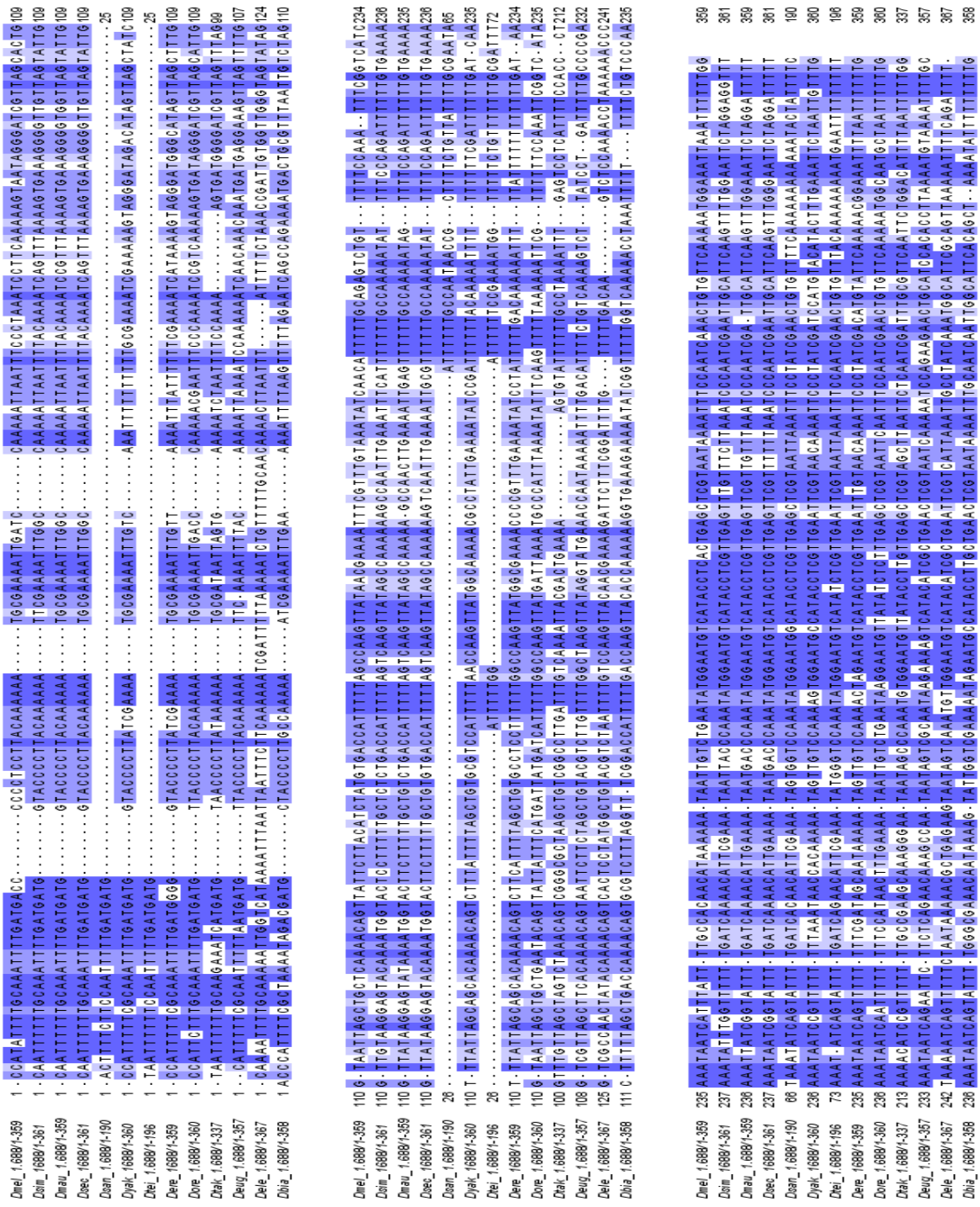

Figure 5. Representative sequence alignment of *1.688* consensus of 13 *Drosophila* species described in this study indicating the presence of conserved regions. Dark blue indicates

regions with higher sequence conservation while white regions indicate no conservation among the sequences.

Figure 6

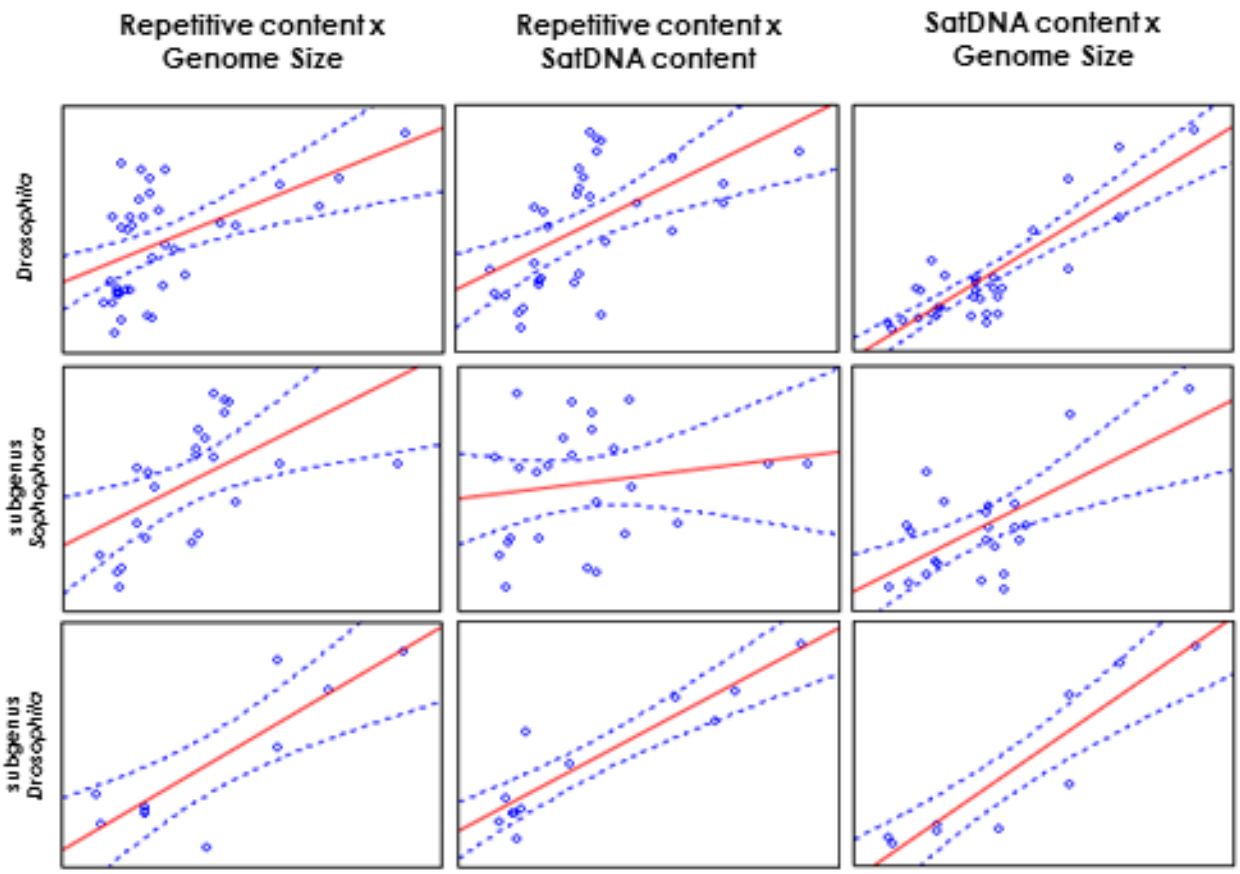

Figure 6. Correlation of repeats with genome size. Graphs show the Spearman's rank correlation between Repetitive Content and Genome Size; Repetitive Content and SatDNA Content; and SatDNA Content and Genome Size in: **(a)** *Drosophila* genus, **(b)** *Sophophora* and **(c)** *Drosophila* subgenera.

List of Tables:

Table 1: Repetitive content estimation and satellite DNA contribution of 36 *Drosophila* species.

Subgenus	Species	Repetitive content	satDNA content	satDNA/ Repetitive content ratio	satDNA families	Genome Size (Mb)
Sophophora	***D. affinis***	34.5	2.07	0.06	4	200.5
	D. ananassae	28.7	3.68	0.13	6	196.6
	D. biarmipes	26.1	1.27	0.05	7	200
	D. bipectinata	32.1	4.72	0.14	7	204.6
	D. burlai	24.8	3.12	0.12	5	N/A
	D. elegans	26.3	4.01	0.15	8	192.2
	D. erecta	8.5	1.62	0.19	3	158.9
	D. eugracilis	25.1	10.89	0.43	3	228.9
	D. ficusphila	14.6	1.76	0.12	2	190.8
	D. kikkawai	19.9	4.85	0.24	3	210.2
	D. leontia	15.2	1.81	0.12	6	N/A
	D. malerkotliana	33.6	6.04	0.18	6	204.9
	D. mauritiana	10.5	4.86	0.46	7	157.9
	D. melanogaster	21.9	6.1	0.28	5	174.5
	D. pseudoobscura	27.1	5.48	0.21	4	193
	D. persimilis	15.7	5.87	0.37	4	193.7
	D. rhopaloa	29.7	4.67	0.16	2	193.9
	D. santomea	24.1	2.75	0.11	5	171.5
	D. sechellia	17.1	7.72	0.45	7	166.7
	D. simulans	11.1	4.53	0.40	8	159.6
	D. subobscura	12.80	1.4	0.11	3	150
	D. takahashii	33.3	3.95	0.12	5	207.3
	D. teissieri	24.4	2.09	0.08	4	166.3
	D. yakuba	15.1	2.84	0.19	7	170.7
Dorsilopha	***D. busckii***	16.3	1.1	0.06	5	139.9
Drosophila	***D. albomicans***	31	16.31	0.52	1	250
	D. americana	27.7	19.75	0.71	7	280
	D. arizonae	13	0.54	0.04	2	142.1
	D. buzzatii	14.3	1.9	0.13	2	169
	D. hydei	10.3	2.16	0.21	5	206.8
	D. mojavensis wrigley	14.9	2.49	0.16	2	170
	D. mojavensis baja	14.2	1.76	0.12	2	170
	D. montana	39	27.41	0.70	5	N/A
	D. novamexicana	21.3	6.82	0.32	7	250
	D. seriema	26.2	2.9	0.11	4	N/A
	D.virilis	32	21.63	0.67	7	325.4

Table 2: Spearman's correlation coefficient among Repetitive DNA content, satDNA content and genome size variation in *Drosophila* genus, and both subgenera *Sophophora* and *Drosophila* independently.

	Subgenus	n	Spearman correlation	p-value	t(N-2)
SatDNA Content / Genome Size	*Drosophila+Sophophora*	33	0,654238	0,000036	4,816477
	Drosophila	10	0,947804	0,000030	8,407647
	Sophophora	23	0,380435	0,073328	1,885117
Repetitive Content /Genome Size	*Drosophila+Sophophora*	33	0,658029	0,000032	4,865586
	Drosophila	10	0,658539	0,038404	2,475105
	Sophophora	23	0,712451	0,000137	4,652637
SatDNA Content /Repetitive Content	*Drosophila+Sophophora*	37	0,513307	0,001158	3,538506
	Drosophila	12	0,867133	0,000260	5,505405
	Sophophora	25	0,205385	0,324673	1,006446

Supplementary Figures:

Figure S1.

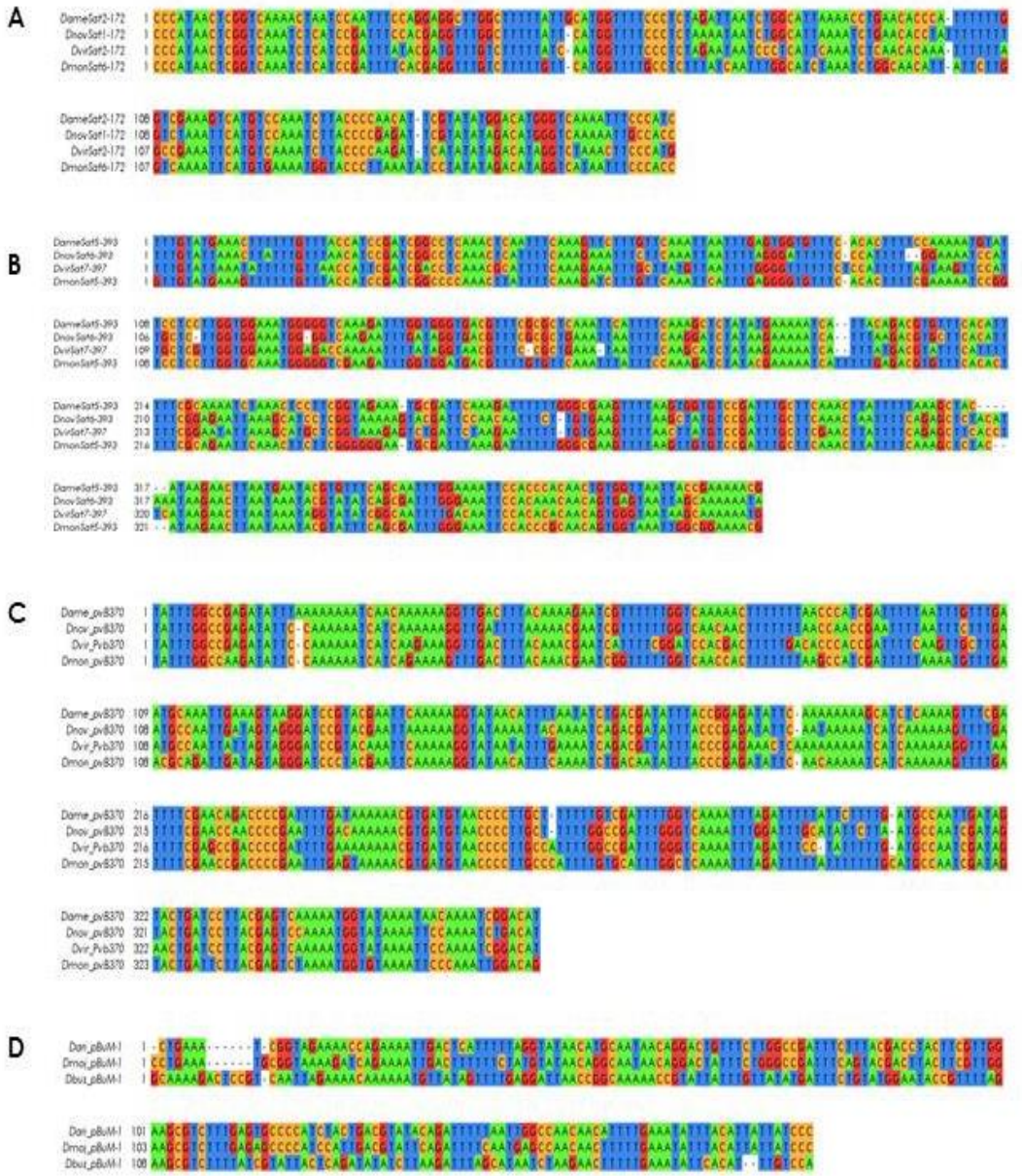

Figure S1. A-C. Sequence alignment of complex satDNA families shared among species from *Drosophila virilis* group. **D.** Sequence alignment of pBuM satDNA family consensus shared among *D. arizonae, D. mojavensis,* and *D. buzzatii* genomes.

Figure S2.

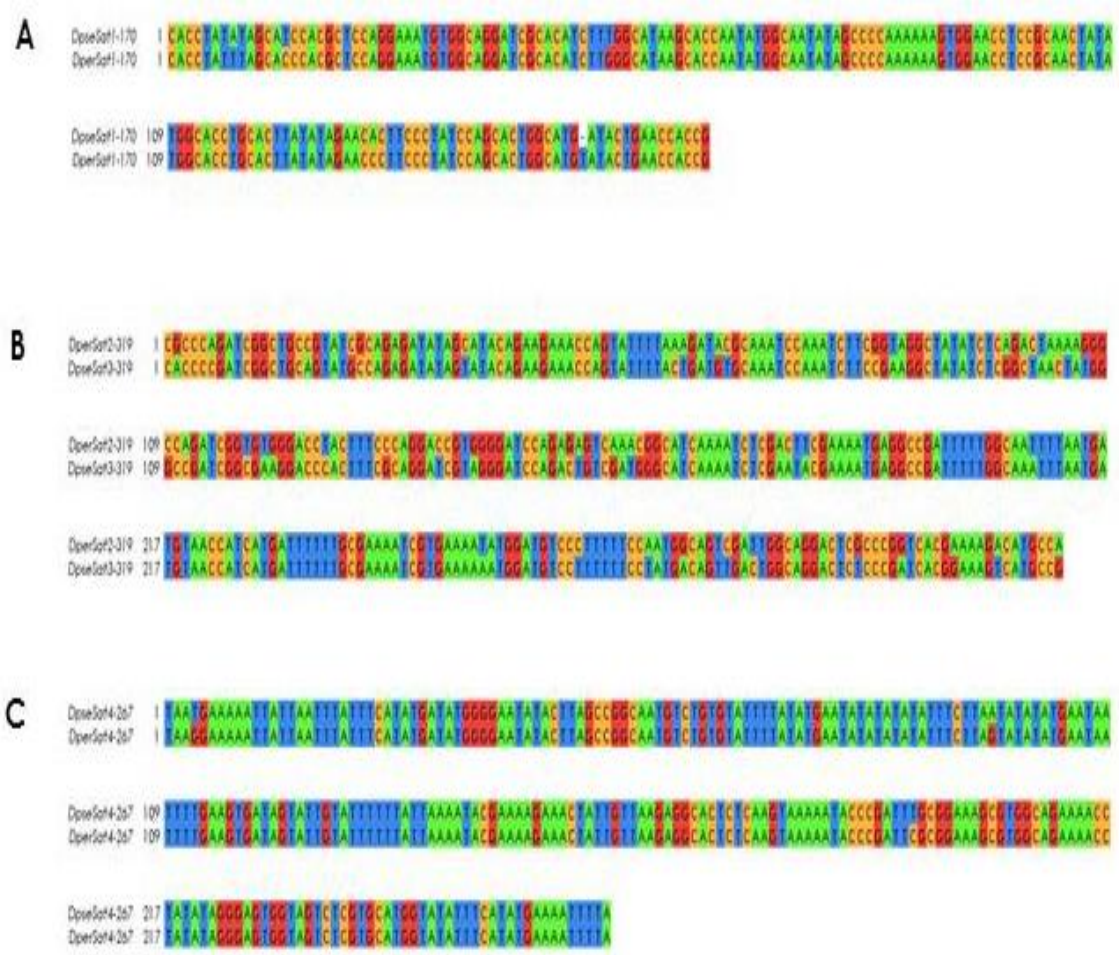

Figure S2. Sequence alignment of complex satDNA families shared between *Drosophila pseudoobscura* and *D. persimlis* genomes.

Figure S3.

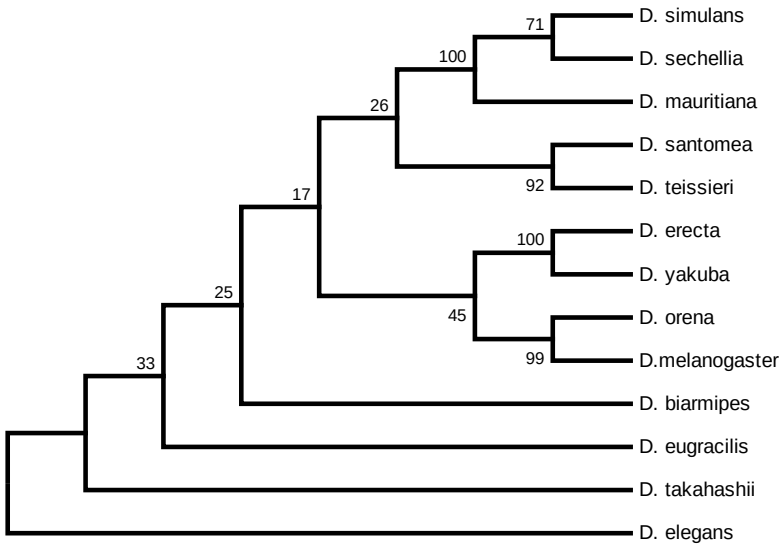

Figure S3. Maximum Likelihood phylogenetic relationship among 13 consensus sequences of *1.688* satellite DNA family on *melanogaster* group. The tree was constructed with T92+G evolutionary model of substitution and 1,000 replicates of bootstrap.

Figure S4.

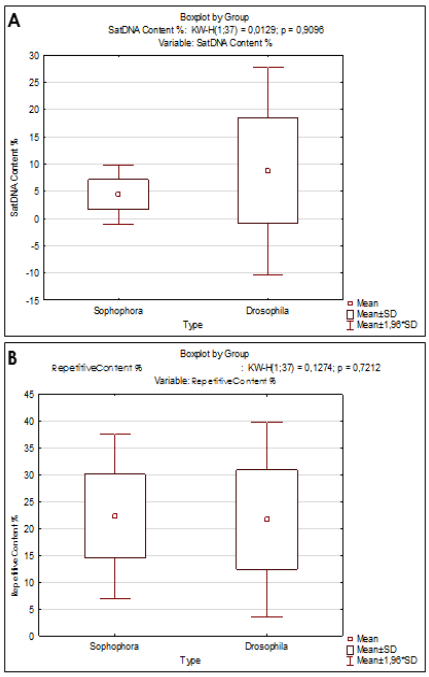

Figure S4. Boxplots of significance values from (a) Repetitive DNA and (b) Satellite DNA contents identified on both subgenera of *Drosophila.*

Supplementary Tables

Table S1. Estimation of genome size used on this study and total number of 100 bp Illumina reads used for each species and its respective genome coverage.

Species	Subgenus	Genome Size (Mb)	Total reads analyzed	Genome Coverage	NCBI/SRA Files
D. affinis	*Sophophora*	200.5	3534015	1,7670	ERX103525
D. albomicans	*Drosophila*	250	2671499	1,0685	SRX010928
D. americana	*Drosophila*	280	1326207	0,4736	SRX2582810
D. ananassae	*Sophophora*	196.6	5072230	2,5878	SRX144727
D. arizonae	*Drosophila*	142.1	5027345	3,5403	SRX1065941
D. biarmipes	*Sophophora*	200	6693021	3,3465	SRX095626
D. bipectinata	*Sophophora*	204.6	1666248	0,8167	SRX094522
D. burlai	*Sophophora*	N/A	1823514	-	SRX883275
D. busckii	*Dorsilopha*	139.9	9243029	6,6021	SRX265057
D. buzzatii	*Drosophila*	169	2165197	1,2811	N/A*
D. elegans	*Sophophora*	192.2	2408863	1,2546	SRX094536
D. erecta	*Sophophora*	158.9	5045489	3,1732	SRX997779
D. eugracilis	*Sophophora*	228.9	2392509	1,0493	SRX095624
D. ficusphila	*Sophophora*	190.8	4884243	2,5571	SRX095448
D. hydei	*Sophophora*	206.8	1000000	0,4835	DRX055253
D. kikkawai	*Sophophora*	210.2	1801891	0,85804	SRX095451
D. leontia	*Sophophora*	N/A	2227008	-	SRX883299
D. malerkotliana	*Sophophora*	204.9	1431137	0,7015	SRX237739
D. mauritiana	*Sophophora*	157.9	2243821	1,4201	SRX183513
D. melanogaster	*Sophophora*	174.5	4974307	2,8587	SRX1961048
D. mojavensis wrigley	*Drosophila*	170	2174346	1,2865	SRX2932915
D. mojavensis baja	*Drosophila*	170	3158945	1,8691	SRX1065937
D. montana	*Drosophila*	N/A	1572497	-	SRX1604922
D. novamexicana	*Drosophila*	250	1000000	0,4	SRX2582808
D. orena	*Sophophora*	280.7	3959570	1,4141	SRX997798
D. pseudoobscura	*Sophophora*	193	1519783	0,7874	SRX204754
D. persimilis	*Sophophora*	193.7	1000000	0,51817	SRR363439
D. rhopaloa	*Sophophora*	193.9	1628575	0,8438	SRX095455
D. santomea	*Sophophora*	171.5	3189067	1,8649	SRX752500
D. sechellia	*Sophophora*	166.7	6918830	4,1679	SRX287396
D. seriema	*Drosophila*	N/A	2144275	-	ERX2037878
D. simulans	*Sophophora*	159.6	1260769	0,7899	SRX1799314
D. subobscura	*Sophophora*	150	6337592	4,2250	DRX055266
D. takahashii	*Sophophora*	207.3	2049860	0,9902	SRX095457
D. teissieri	*Sophophora*	166.3	3929041	2,36689	SRX854063
D.virilis	*Drosophila*	325.4	1000000	0,3076	SRX669289
D. yakuba	*Sophophora*	170.7	3829418	2,2525	SRX494771

*Genome sequences were retrieved from *Drosophila buzzatii* genome Project webpage (http://dbuz.uab.cat/welcome.php).

Table S2. Description of all satDNA families identified by *RepeatExplorer* pipeline in the present study from 36 species of *Drosophila*.

	satDNA family	Length (bp)	GC Content (%)	*Clusters Number*	Genomic contribution (%)
D. affinis	***DaffSat1-149***	149	51.67	2	0.911
	DaffSat2-131	131	29	2	0.678
	DaffSat3-342	342	21.63	1	0.328
	DaffSat4-253	253	26.48	1	0.154
D. albomicans	***DalbSat1-35***	35	11.76	2	16.31
D. americana	***DameSat1-7***	7	14.28	1	9.501
	DameSat2-172	172	40.1	3	4.126
	pvB370	370	29.18	4	3.048
	DameSat4-7	7	28.57	1	1.950
	DameSat5-393	393	35.29	1	0.177
	Tetris-220	220	29.8	1	0.173
	DameSat7-32	32	37.5	1	0.123
	DameSat8-36	36	50	1	0.069
D. ananassae	***DanaSat1-35***	35	40	4	1.41
	DanaSat2-180	189	36.4	2	1.06
	DanaSat3-7	7	42.8	1	0.364
	DanaSat4-201	201	41.1	2	0.544
	DanaSat5-273	273	41.3	1	0.157
	DanaSat6-25	25	68	1	0.149
D. arizonae	***DariSat1-130***	130	28.2	1	0.348
	pBuM	185	37.5	1	0.192
D. biarmipes	***DbiaSat1-189***	189	36.6	1	0.310
	DbiaSat2-263	263	32.3	1	0.203
	DbiaSat3-190	190	36.5	1	0.152
	DbiaSat4-132	132	28.7	2	0.313
	DbiaSat5-209	209	35.4	2	0.118
	1.688	360	34.9	1	0.091

	DbiaSat7-201	201	31.6	1	0.090
D. bipecnata	***DbipSat1-150***	150	28	1	1.310
	DbipSat2-191	191	38.6	1	0.958
	DbipSat3-189	189	38.9	3	1.819
	DbipSat4-205	205	29.1	1	0.362
	DbipSat5-146	146	22.6	1	0.204
	DbipSat6-190	190	38.9	1	0.033
	DbipSat7-142	142	64.2	1	0.033
D. burlai	***DburSat1-135***	135	30.3	1	1.860
	DburSat2-300	300	49.7	1	0.349
	DburSat3-9	9	22.2	1	0.341
	DburSat4-45	45	45.4	2	0.386
	DburSat5-370	370	18.7	1	0.182
D. busckii	***DbusSat1-11***	11	9.1	1	0.503
	DbusSat2-550	550	31.8	1	0.230
	DbusSat3-167	167	23.3	1	0.148
	DbusSat4-29	29	51.7	1	0.120
	DbusSat5-105	105	29.5	1	0.091
D.buzzatii	***pBuM***	189	29	1	1.71
	CDSTR198	198	34	1	0.23
D. elegans	***DeleSat1-150***	150	27.9	1	1.390
	DeleSat2-10	10	10	1	1.240
	DeleSat3-272	272	27.5	2	0.579
	1.688	374	29.7	2	0.53
	DeleSat5-190	190	29.14	2	0.210
	DeleSat6-160	160	35	1	0.058
D. erecta	***DereSat1-250***	250	34.6	2	1.138
	DereSat2-181	181	29.8	1	0.309
	1.688	360/191	30.2	1	0.172
D. eugracilis	***DeugSat1-8***	8	50	5	5.691
	Deug_1.688	360	30.4	4	4.088

	DeugSat3-112	112	34.57	1	1.110
D. ficusphila	***DficSat1-197***	197	32.32	3	1.682
	DficSat2-189	189	39.26	1	0.08
D. hydei	***DhydSat1-7***	7	42.8	1	0.733
	DhydSat2-6	6	33.3	2	0.473
	DhydSat3-14	14	28.5	1	0.386
	DhydSat4-2	2	50	1	0.294
	DhydSat5-37	37	32.4	1	0.282
D. kikkawai	***DkikSat1-41***	41	34.14	2	2.493
	DkikSat2-109	109	27.52	1	2.180
	DkikSat3-19	19	47.36	1	0.177
D. leontia	***DleoSat1-41***	41	34.14	1	1.340
	DleoSat2-19	19	42.1	1	0.176
	DleoSat3-111	111	67.5	1	0.146
	DleoSat4-109	109	28.4	1	0.064
	DleoSat5-28	28	35.7	1	0.058
	DleoSat6-570	570	40.7	1	0.026
D. malekortliana	***DmalSat1-188***	188	35.1	2	2.405
	DmalSat2-32	32	71.8	1	1.940
	DmalSat3-197	197	41.1	1	0.621
	DmalSat4-191	191	37.6	1	0.569
	DmalSat5-205	205	33.3	1	0.415
	DmalSat6-380	380	54.7	1	0.088
D. mauritiana	***DmauSat1-15***	15	40	3	3.586
	Dodeca	11-12	58.3	1	0.458
	1.688	357	30.08	1	0.460
	DmauSat4-147	147	31.9	1	0.200
	DmauSat5-197	197	31.97	1	0.117
	Responder	120	38.9	1	0.028
	DmauSat7-39	39	48.7	1	0.010
D. melanogaster	***TGTTATTCTA***	10	20	1	1.750
	AGAGA	5	40	2	1.199

	1.688	360/353 /260	30.91	6	2.01
	Responder	120/240	35.34	4	0.764
	dodeca	11-12	66.66	1	0.380
D. mojavensis wrigley	***CDSTR130***	130	26.2	1	1.63
	pBuM	185	26.5	1	0.86
D. mojavensis baja	***CDSTR130***	130	26.2	1	1.060
	pBuM	185	26.5	1	0.703
D. montana	***DmonSat1-3***	3	66	1	19.70
	DmonSat2-4	4	50	4	4.33
	DmonSat3-55	55	50	1	1.94
	pvB370	370	32.88	1	0.621
	DmonSat5-393	393	31.80	1	0.508
	DmonSat6-172	172	36.04	1	0.302
D. novamexicana	***DnovSat1-172***	171	33.9	2	3.031
	pvB370	369	31.4	1	1.180
	DnovSat3-6	6	50	1	0.789
	DnovSat4-190	190	29.7	1	0.720
	DnovSat5-33	33	42.4	1	0.688
	DnovSat6-393	393	31.8	1	0.248
	DnovSat7-100	100	31	1	0.165
D. orena	***DoreSat1-172***	172	24.5	1	10.400
	1.688	360	28.8	3	1.312
	DoreSat3-181	181	28.7	1	0.603
D. pseudoobscura	***DpseSat1-170***	170	47	1	1.930
	DpseSat2-21	21	50	1	1.680
	DpseSat3-310	310	44	2	1.186
	DpseSat4-267	267	27	1	0.689
D. persimilis	***DperSat1-170***	170	47	1	5.2
	DpseSat2-319	319	50	1	0.495
	DpseSat3-21	21	44	2	0.162

	DpseSat4-267	267	27	1	0.020
D. rhopaloa	***DrhoSat1-191***	191	33.1	3	4.345
	DrhoSat2-11	11	20	1	0.331
D. santomea	***DsanSat1-9***	9	33.3	1	1.82
	1.688	191	29.47	2	0.462
	DsanSat3-396	396	35.8	1	0.309
	DsanSat4-96	96	34.37	1	0.090
	Responder	165	40	1	0.051
	DsanSat7-241	241	45.22	1	0.020
D. sechellia	***DsecSat1-15***	15	40	1	6.040
	Responder	120	39	1	0.860
	DsecSat3-5	5	40	1	0.274
	Dodeca	11-12	66.6	1	0.265
	1.688	361	30.05	1	0.201
	DsecSat6-193	193	36.2	1	0.077
	DsecSat7-39	39	53.8	1	0.005
D. seriema	***pBuM-2***	370	23.9	1	1.93
	DBC-150	150	55.9	1	0.81
	CDSTR138	138	31.2	1	0.22
	CDSTR198	198	34.8	1	0.02
D. simulans	***DsimSat1-15***	15	40	1	3.180
	Dodeca	11-12	66.6	1	0.583
	Responder	120	38.4	1	0.463
	1.688	360	29.4	1	0.120
	Dsim5-135	135	33.3	1	0.103
	DsimSat6-193	193	31.6	1	0.041
	DsimSat7-241	241	64.5	1	0.033
	DsimSat8-39	39	48.7	1	0.012
D. takahashii	***DtakSat1-190***	190	34.21	1	0.987
	1.688	336	35.0	3	1.537
	DtakSat3-33	33	42.4	3	0.922
	DtakSat4-307	307	26.7	1	0.390

	DtakSat5-132	132	40.4	1	0.122
D. teissieri	***DteiSat1-10***	10	20	1	0.572
	DteiSat2-114	114	23.68	2	0.599
	DteiSat3-132	132	34.84	1	0.351
	1.688	191	25.12	2	0.567
D. virilis	***DvirSat1-7***	7	28.57	1	15.900
	DNAREP-TR1	154	30.16	1	1.550
	DvirSat2-172	172	35.46	1	0.973
	pvB370	370	34.59	4	1.753
	DvirSat4-7	7	42.8	1	0.187
	DvirSat5-132	132	45.45	1	0.169
	DvirSat6-397	397	30.80	1	0,115
	Tetris-220	221	29.86	1	0.081
D. yakuba	***DyakSat1-132***	132	30.5	1	0.599
	DyakSat2-320	320	25.3	3	1.184
	DyakSat3-396	396	36.11	1	0.439
	1.688	360/191	25.64	2	0.581
	DyakSat5-434	434	44.47	1	0.019
	DyakSat6-62	62	48.38	1	0.015
	Responder	120	39.39	1	0.010

Table S3. Interespecific genetic distances (*p*-distance) between *1.688* satellite DNA consensus sequences of 13 species of *Drosophila melanogaster* group.

D. melanogaster												
D. simulans	0,331											
D. mauritiana	0,305	0,075										
D. sechellia	0,304	0,061	0,031									
D. santomea	0,335	0,326	0,275	0,289								
D. yakuba	0,321	0,333	0,324	0,319	0,344							
D. teissieri	0,284	0,291	0,246	0,255	0,170	0,323						
D. erecta	0,311	0,337	0,331	0,323	0,330	0,142	0,299					
D. orena	0,215	0,331	0,296	0,297	0,317	0,308	0,277	0,315				
D. takahashii	0,365	0,351	0,325	0,330	0,399	0,369	0,376	0,342	0,357			
D. eugracilis	0,369	0,375	0,372	0,361	0,380	0,362	0,342	0,358	0,351	0,377		
D. elegans	0,383	0,390	0,374	0,366	0,387	0,440	0,388	0,430	0,434	0,401	0,409	
D. biarmipes	0,384	0,359	0,341	0,333	0,342	0,351	0,321	0,352	0,391	0,374	0,384	0,415

Supplementary Material 1. Nucleotide sequence consensus from all satDNA families identified in the present study. Nomenclature of all new satDNA families were done according to Ruiz-Ruano et al. (2016).

>DaffSat1-149
TTTCTAGGGCCAGGAATTTCACGCAGAACCAATTCCTTAGCCCCCCATTGATAGTGGCGGCCATTTCCCAATTTA
CCCACTGGCAATGGCGGCCATTGCTCACACCACCGTCATTTGTACCCAGTGAAGCCTGGGCTATCACTTACACA
>DaffSat2-131
TAACAAATAATATCCAGAGAAAACTATGTTGAAATAGGTTTTTATTGATTTTTAATATTTTGGACTCGGTCAACC
TTTTGCGAAAAATGCGCATATGTCCTTTAATTAAGTTGGCTGAGTTTCAATATCTT
>DaffSat3-342
TACATAGAAAATAAACTTGAGGATGATATGAAGTTTATTAATGGAGGAGATATAAAGACATATATCATATGCATA
TGTTGAATTATGATTTTATAAAATAAATGAATATAACAAATAACTTGGAGTGGTGCCCATATAAGATATTGTAAT
ATGTACGGAATTAATAAATATATATATAATATAATGAAATATTCATTATAATGAGACCGTCATGTATAAACTACA
AATAAGAATTGAACGATGTTTTGATATTATATAAAAATTGATTTTGTATTTAGAATATCAATATTATGGTTTAAT
ATGGATATGGGAATATTTCCCATATAACTATATGATATACAA
>DaffSat4-262
TTATACTTGAGAGTTCCATTTCAGTAATAATTCGGTCTCGTATTTTAATGCAAGAAACTTTATTATCACATTAGA
AATATTCATACATATGTACATATTAACAAATATATATCCATATAAAATATCAGACATTTACGGTAAGTATCATAT
GAAATGAATTATTTCTTATTATATTTATTTTTTTCATATGACATATACCAAGCGCGACACTCCATATATAGGTTT
TTTGACACGGCTTCTGAAAATCGGTCAT
>DalbSat1-35
AAAATTAAAAAATTGTAAAAATCTGCAAAATTTT
>DameSat1-7
CTACAAA
>DameSat2-172
ATATGGACATGGGTCAAAATTTCCCATCCCCATAACTCGGTCAAAACTAATCCAATTTCCAGGAGGCTTGGCTTT
TTATTGCATGGTTTTCCCTCTAGATTAATCTGGCATTAAAACCTGAACACCCATTTTTTGGTCGAAAGTCATGTC
CAAATCTTACCCCAACATTCGT
>Dame_*pvB370*
GATCCTTACGAGTCAAAAATGGTATAAAATAACAAAATCGGACATTATTTGGCCGAGATATTTAAAAAAAATCAA
CAAAAAAGGTTGACTTTACAAAAGAATCGTTTTTTGGTCAAAAACTTTTTTTAACCCATCGATTTTTAATTTGTT
TGAATGCAAATTGAAAGTAAGGATCCGTACGAATTCAAAAAGGTATAACATTTTAATATCTGACGATATTTACCG
GAGATATTCAAAAAAAAGCATCTCAAAAGTTTCGATTTTCGAACAGACCCCGATTTTGATAAAAAACGTGATGTA
ACCCCTTGCTTTTTTGTCGATTTTGGTCAAAATTTAGATTTTTATTCTTTTGATGCCAATTGATAGTACT
>DameSat4-7
CAAACAA
>DameSat5-393
GTTTGTATGAAACTTTTTTGTTTACCATCCGATCGGCCTCAAACTCAATTTCAAAGTTCTTTGTTCAAATTAATT
TGAGTGGTGTTTCACACTTTTCCAAAAATGTATTCCTCCTTGGTGGAAATGGGGGTCAAAGATTTGGTGGGTGAC
GTTTCGCGCTCAAATTCATTTTCAAAGCTCTATATGAAAAATCATTACAGACGTGTTTCACATTTTTCGCAAAAT
CTAAACTCCTTCGGTAGAAATGCGATTCAAAGATTTTTTGGGCGAAGTTTTAAGTGGTGTCCGATTTGCTTCAAA
CTTATTTTTAAAGCTACATAAGAACTTAATGAATACGTGTTTCAGCAATTTGGAAAATTCCACCCACAACTGTGG
TTAATTACCGAAAAAC
>Dame_Tetris-220
ATCATACTGTTAAATGGCAGCCATATTTTAAACTTTATTATTTAATTTCTCCGAAAACGGCCCCAAATGCACGGT
AGTTGTGCATATAGAGCACGATTTATCAATCATTTTGCTATTTATTTCATTAAAATCATCGGTACAGATTTTGAG
AAATTCTCATTTTTCTATTTCTCTCATAACATCCTATGTTAATGCAAAAAAACAAGCTGGCAATATTATAT
>DameSat7-32

TGACAAAAGCTGATTGCTATATGTGCAATAGC
>DameSat8-36
TCCAAAACGACATAACTCCGCGCGGAGATATGACGT
>DanaSat1-35
TTTTCTAGCGGTTTTTAGCGCTAAATCAGCGATGA
>DanaSat2-180
AGGAAATTTGACAAAAAATCTAAAAAAAAAAATAGTTTCCCTGATTTTGATGCAGAATGGCGGAGAATCGATCTAC
AAATCGTTTAAGGTATTCCGCTAGAGAATCGGATGATTTTTGGCAAAGATATAGCCTTGGCAGTGGGTCATATCG
GGCACTCCTGCGTTTTCCAACATTTTCAAGGGGATACCCC
>DanaSat3-7
TTGACCT
>DanaSat4-201
AAACGGCGTACCATTTATGAATCAGGGAACTAATCCTCGTCAATCTGCATCAAAGGATCCTTAAAATTCTCAAAA
TTCAATTTTGGCCCCATTTTTGGCCCAAATCATGATGGTACCCCTTTGGAAAACCGTTGGAAAAGTGCGATCAGC
ACTTTCCTCGCAGTTTCTGGCGAATATCTTTACTATTTCGCATCCGATTGAA
>DanaSat5-273
CGCGTTACATTACTTGTATATTCGGTACCGAGTGGGTCGGCCACTGCATTTTGAGGGGTTAAACTGAAATTGTAA
TGGAAATTGGCCGCTATTAAACATTTACTTGCGTTAGAAACAGTATAACTTACTTTAGAATTTTTGGCGGGGTAA
GGTTACCGAAGCGGTCCTCGTACCCGCAACAACAAATTTACCACCCAAATTCTAGTCGATATGGTGAATATATTG
CAGTAAATCCCTCGCACACGCATTTACATCTAGTGTTGAACTCAGCAT
>DanaSat6-25
GCCAGCCAGGTCACCTCCTCCAGCA
>DariSat1-130
CAATAGAGAAAAATTAATGCAGTAATTATTGCAACATGTGATGGAGAATCAGAAAAAGCAAAGCGATTTAATATT
TTGAAAATATAACAGAAAATTTTAAATTATTCGGCAATAAATTATTAAGTATTG
>pBuM_Dari
CTGAAATCGGTAGAAAACCAGAAAATTGACTCATTTTTAGGTATAACATGCAATAACAGGACTGTTTCTTGGCCG
ATTTCTTTACGACCTACTTCGTTGGAAGCGTCTTTGAGTGCCCCATCTACTGACGTATACAGATTTTTAATTGGC
CAACAACATTTTGAAATATTTACATTATTATCCC
>DbiaSat1-189
GAAAAGTTTAAATGTAGAATATTAGGAAATTCAAACACAAACACAAAGAGGTATCACACTTCAAAATCGGAGGCC
TATAACCCAAGTTATGGAATGCTGAAGTGCAGTTTTGGGTCACCATACTAAAAACCATTGGAATTACGGAAAAAG
CGAACATGAAAAGGGCTTAAAATTTCGACCCTCCGTAAAAG
>DbiaSat2-263
AATTTTCGCATCAAGAAAGGTTTGACCACACTGTCTGCAGTGTAAAATTAAAATCTTCTTTAAAAGTACAATCTC
GGTAAATTCTACATGTTTCTTCTTTGGGAGATTGTTGGTCAAAATATCAGCAATCATTTCATTAGTACATTTATA
CTTTAACAAAACTTAACCTTCCTAGACAACTTGTCTTACAAAATGATAGCGTATATCAATATGTTTTCTTCTAGA
ATGATGAACTGGATTCTTAGCCAGAGCTTGGGAACTCA
>DbiaSat3-190
CATGATTTTTTGGATTTCAGATTTTAGCAAAAAAAACTCCTTTTCTCGCGGTTTTTCAAACGTAGTTTTTCCAAA
TCAGGGCGCACAGCAAAAAGACCGACTGTTTTGGAAAGCTTGCTTAATTTGCTAACGATCTGCATCATTTCTGGG
AGAATTTATTTTTTGACCCATTTTCGCATTTTTTTTAGGGGTCCCAT
>DbiaSat4-132
AAGGTCTAGATTTTCAAGTAAGAGACCAATTTGATTAACATATCAATATAAGTACAACTCGACCATCTAATAATA
ACAATATGAATCGTCATCTTATTAGTGACGCGAAGATTGGCAAACATATAATATAAA
>DbiaSat5-209
AATTAAAGGTAGAGTTTGCCAACATGCTGGAAGAAACTAATTTGCAAAAAGATGAAAGCTAAGGGACATTCATAG
CAAGATTGATTCGGAGGTGACCCATCCAAGAACTTCACAGCGTTTGTGACGCACAACAATTTGCTAGAGCGAATT
AATATTTTACAAGTGCAACAAGAAAGAACACATATATAAAAATTGAGATGAGAATAATC
>Dbia_1.688
ACCTTTTTTGGTGTAACTTGGTCAAAAATGGTCCGAAACCTAAAGACGCACTGTTTTGGTCAGCTAAAAAGCTAG
ACAATTAAACGCAGTCATTTTCTGGCTGATTCTAAAAACTTAAAATTTTTTCAAATTTTCGATTTTTTGGCAAGG

GGTAGCATCGTCTATTTTAGCGAAAATGGGTCAAAAAAAAATATTTTAGCTGTGAATGCCAATTGATTGCAAATT
TTATTACGAGTTCAGCAAGGTATGGCATTCTATTTTTCGACCAATACTTTTATTGTTTCGCCCAAAATACTGATT
ATTTTTGGGACAGAAAAAAAAATTTAGGTTTTTGACCAAAAACCGATATTTTTCTTTC
>DbiaSat7-201
AAAAATAAATTGTCCCCAAAGTGACGAGGATCGTTAGCATTTTCAGCACACATTTCAAAACAGTCGGTCTTTTAG
CTGTACGACTTCATTTGACTAAACTAAAAACAAAAAACCACTAAAAATAAGAATTTTTTGACAAAATCTGAGTTT
AAAAAAATGACGACCCCCACCCTTACAAAAAATGCGAAAATTCTTCA
>DbipSat1-150
AAAAGGAGGAAACACAATGTCACAATGTTGAACATACAGTTTTTAACATATTTAATAAACTTCATTTTACAATAG
ACTTGACTGTGATACATTAATTTAACAAGAAAAAGGGACATTTTATAATTTGGAGAAAAAAGGAGGAAAAGGAGG
>DbipSat2-191
TAAAAACCATATTAAGGCAATGTTTTGAATGAGGAAAGTAGCTTTTGAGTTCCCTGATTACAAAACGGTAGTTTA
TTTCCTGATCGGATACAAATTGGAAGAGAGACGGCCACCGGAAGGGGAAAAAGTGTCGAAATTGCGAAAACCCAC
AATTTTGCAGGGGTACCCCTAGGAAAAAATTCGAAAAATT
>DbipSat3-189
CTACAAATCGTTTAAGGTATTCCGCTCAAGGATCGGATAAGAAATGGCAAAGCTATAGCCATCGCTGTGGACCAT
ATGAGTCCTCCCCATACATTCCCCAATTTTCAGAGGATTCTCACCAGAAAAATGACAAAAAATCTAAAAATATTT
TTGTGTTGGGGTTATAATGCAGTTAAGTAGAGAATCAAG
>DbipSat4-205
CTTGCAAATAAATGTAAGACACGAAAACACTAAAATAAGGTTGTTTAGGCTGACTCTTTTTTTAATAATTTTTCT
GGGTCTTTTGGAACAAGCAGATAAACTGAAAAATAAACGAATGAGTGCGCCATCTTTTGTTCACTTTTTAATAGA
AATCAAATTATTGTTTCGTCTCGATACTAACATAACAATAAATTCTTGTGCGTAA
>DbipSat5-146
GCGCCGTCCAAAAAACCTATACAAACAATCAAAACATAAAAAAATACAAAAAAAATTTACAGAATTTTTTTTGA
TGTAAACATGAAAATATACAAACACTAATTTAACGTATTTTTTTTATGTTTCAATGTTAAAAAACGCTCTA
>DbipSat6-190
AAAAACGGAATACCATTTATGAATCGTAGATCCTTTGCCCGCCAATCTGCATCCCAAGATCTGTAAAAAACCCAA
AATAAAAAAATTGGCCCAGGTTCTGGCCAAAATCTTGATGGTACCCCTTGGAAAAGTGATAATTCTTGCTGTTGGC
CAGTTTTTAGTGAATATCTTAACTATTTCGCAACCGATTC
>DbipSat7-142
GAGGCGTGGAGTGCCCAGCCGATTCCCCTACCGACCCGTCGAGTGCCTAGCCGATCCCCTGCCAGCTGATAACCC
GCCAAGAAATGAAGTGCCTGCCCAATCTCCCTGCCAAAGCCTGGAGTGCAAGGCCGATCCCCGTGCC
>DburSat1-135
TAGATACAGAACGAAATAGCACAGAAAAGCTGTCGGAATCAAAATTAAAGCATTTTTGACGGAATTGTAACTCCA
AAAAAAAAAACGAACAAAAATTAAACTTCTGAAATCTGGAACTTTAAGTATGCAATTGCA
>DburSat2-300
CGCCTTTTAGCAGTACAAAATCATCAAATGCTGGATTTCATGCACCAGGCGCGCAGTAACAGCAGCAAATAGATA
ATAAAAAAATTCCACCCCTTAGTGTGACCACATTGGCAACTTTGTTAGGTGTGACCATATTGGCAAACGTCGTAT
TTGTCTATCGATTATCGATCTGGCGTTGCCAGACTTTTCGAACGCAAGCCGATGGTGCTGCCAGACTTTTAACTT
TGTGTGCCCAGCTGCCGTGGAGGCGGAAATGTTTTTGTTGTATTGCCAATGTTGTTGTAGGCAAAAACAAGGTCT
TAATAC
>DburSat3-9
TCGTATTTT
>DburSat4-45
GACAAAATGGCATTCTTACCGCCATCGCTGTGTTGCAATATTGG
>DburSat5-370
TATTAAAAGACATGATATAAAGATTATTATGATAATATATAGAATATATTGAGAAAATAAAGATGGAATATTACT
GATAAATATATATGTGGAAATATATCATTGTGAAAATAAAGTTGGAAAATATATGAAAATATCATGTTTAATATT
ATTGATTCTTCTCATAATGTATATTTGATTATATACTAACATATTGAAAAAAACTTTTTATATAGAAAAATGCCA

TGAACTGATTATATATTGAAATAAAATAAACATTGAGAATATCAAGTTGAAAATATAAAGTCCATTTCTTGGACT
TAGAAAATATATGAAATTAACATGTTTGATATATAACAATATAACATTAAGAAAGAACTCCA
>DbusSat1-11
TTTTTGAACAT
>DbusSat2-550
ATTTCGCTGAGTTGCGCCTTATCTAAATGCGATATTTCAGAAAAATTTTTCTATTTTGTTTACTATTTTGTTCTT
AACTTCGCCGAATAAATTTTAAATCATATTTATGCTTATTGTAATCGTTAATGCCACGCCCACATAAAATGGTAA
AAAACATGCCAATAGGATGTGTTTTTTGTGAGTTATAGCTTCGGAAAGCTTAAAATTAACTCTTCAGAATTGGTT
TTATTGCCATAACTATTTCAAAAACAATTTTTTGAACACAACAGCACATGTTTATTTAGTATCTCATGAGTAAAT
TAAGCTTGAAAACACTTCTGTGCGATTTTTTTTGGCTGAGATATAGCTTCTCAAAGTTAGAAAGTGATCAGAAAT
GAAGACGGCATCATTTTTGGCAAAAATTTACAGGGGTTACGATCACGATTTTTGCCAAAATATCGAAAATCGTAG
ACTTTCTAAATTTGAATGCAGTTTTATTCTTATGCATCTCACGAGTCTAACAAGGCATGAAACTCCGC
>DbusSat3-167
TTTGCTTTTATTCGTTTCAAAACGCAAAAAAAACATTTCTTGAAAAAAAAATTTTTTTTCTTCACTTCATATGAA
ATCACATTATTATCAGTACAACAGTCAACCTGATATCCATAGTCTATAGACAAAAAAGTCAAAAAATTGTTGGTA
AAAAATTTTTTTTTTTC
>DbusSat4-29
TCCAATAATCAATGGGTGCTGCCGTGCAG
>DbusSat5-105
ATATGGACCCAAATTTGGCGAAATTTTCATACAAAATACTTGAAAATCTTTAAATGCCCATAACTTCTTAACGGT
GAGAGATAGAGCTTAAATACGATTTAATTT
>Dbuz_pBuM
GCAAAAGACTCCGTCAATTAGAAAACAAAAAATGTTATAGTTTTGAGGATTAACCGGCAAAAACCGTATTATTTG
TTATATGATTTCTGTATGGAATACCGTTTTAGAAGCGTCTTTTATCGTATTACTCAGATATATCTTAAGATTTAG
CATAATCTAAGAACTTTTTGAAATATTCACATTTGTCCA
>Dbuz_CDSTR198
AAGGTAGAAAGGTAGTTGGTGAGATAAACCAGAAAAAGAGCTAAAAACGGCTAAAAACGGCTAGAAAATAGCCAG
AAAGGTAGATTGAACATTAATGGGCAAATGGATGGATAAATAAGACTGGTCATCATCCAATGAACAGAATCATGA
TTAAGAGATAGAAATATGATTAGAAAGTAGGATAGAAAGGTTAGAAAG
>DeleSat1-150
AAGCTGTTGACCTGAATCTTTTCAAATCGTATTTTTTAATATATATCATAATTAATATGAAATATCATGCCAATA
GGAAGTCTATATCTTATACAAAACTCAAATTGCGTTCAAATTAAATTGTTCAGCAAATTTCGCGCAGTGTAC
>DeleSat2-10
AAATATACTT
>DeleSat3-272
GCAACCAAATATATGGGTATTTCTCAGAACATATAAAATTAATATTATATTGTGAATACCTATAATAAATAACAT
TCATATATATGAAAATGTTGTAATATATCAATATTATTGATATATTATATATGAAATGTATGGAAAATATAAAAT
ATTCCTTCAATTTCCTCAAGCATATGGAGGAAATTCATATGAAGGCAGACTTTTAGTACTGTAAACACGCAGTTT
AAAGACGGCGTGTCCAAAAACATATATAGGGTAGTGCCTGTTGGCTG
>Dele_1.688
TCAAAATTTTGCAAAAATTGGTCATAAAATTTAATTAATTTCTTCAAAAAATCGATTTTTAAAAATCTGTTTTTG
CAACAAAACTTAATTTATTTTTCTAAACCGATTGTGGTTGGGTAGTATAGGTCGCCAACTATACAAAACAGTCAC
TCCTATGGCTGTACGTCTAATTTTGTCCAAGTTACAACGAAAAGATTCTTTCGGATTTTGTTTTTGACAAGTCTC
CAAAACCTAAAAAAACCCTAAAAATCAGTTTTTCTAATAAAAACGCTGAGAAGTAATAGTCAAATGTTGGAATGTC
ATACATCGCTGAATTCGTCATTAAATTGCCTATCAAATGGCATTCGCAGTTAAAAATTTTCAGATTT
>DeleSat5-190
TTTTTGTCATTTTTATGATGTAACCCCTTTCTAAATTTTCAAAAAAATAGCTTTGCAGAAAACTGACTAAACACC
TCCAGTTTTAATTCCATAACTTTTCTTACAGACATTCGATTTGAATGTGGTATATCTTTAAAGTCTTGTGGCTAA
AGTATCTAAAAATCTGCATTTAAATCTTGGGTTGTAGTTTCGTTTCAAA
>DeleSat6-160

TGATCATAGTATGCATTAAAAAATACGATTTGGAAAGATTCAGGTCAACAGCTTATAAACTGCCCGAAATATGGT
GAAAAAATTAAATTTGAACGCATTTTGCCCTATTTTTCCAATGGGAGCTATAAGAAATAGACGTCCGATTGGCAC
GCTCTTTCAT
>DereSat1-
250ATTTAGTTAATAAATGTGTTCATGTTTGTGTTTGCGCACGAAAAGTGGTTTCATGTGGTGCGCAGATAAACA
ATCTACATCCAGAAAGAAGAAAATATAAACTCTAAACTCTAGACCAAGTCATCGGTAATTGTAATTAAAAACTGG
TGCACATAGTGTTCAAAAATAATTCCCAACTTGTTATGGCTTATATTTCATTATACGTTCCCTCTAACAGCCTAT
AAAGTAGTGGACAGGAAGTTCCGTGA
>DereSat2-181
AGAAACATTTAGAGGTATGCCACTTCAAAATCCGATTAAGATAAGTTGAGTTATGGACTTTAGAACTTCAGTCTT
GGGACATTTTTGAATTTCTACAACATCAGAACATCAGAAAAATGACAAAAAAAATGTAGAAAAAATAAATTTCCAA
ATTTGAATGCAGATTTTTAGAGAATTGCTCC
>Dere_1.688_360bp
TCACTTTTTGACGGATTTTGGAAAATTAATTTTTCGGCCAATTTTCGCATTTTTTGTAAGGGGTAACATCATCAA
AATTTGCAAAAAATTGCAAAAAATTAGCGTTTCCATTTTTGAACGCAGTTCGATTGGAAATTTAATTACGAGCTC
AACGAGGTATAACATTCCATATTTAGACAATTATTTTTCAAGTTGTGGCCAAAAAACAGATTATTTTATGACCGA
AATTCGAAAAAACGGATTTTGGCAAAAATGCAAATATTTTTGATGGGCATTTTAATCATAACTTGGCTAAAAATG
ATCATAAAGAAAAAAGAATAACTGTTTTGAGCAGCTAATTACCAATACTAACGACCCCTA
>Dere_1.688_191bp
CGAATATTTGCCAAAATGCGTTTTTCCGATTTTCCGACATAAAATAATCAGTTTTTTGGCCACAACTTTAAAAAT
AATTGTCTGAATATGGAATGCCATACCTCGTTGAGCTCGTATTCAAATTCCCAATCGAACTGTGTCAAAAAATGG
AAATTATATTTTTTTGCCATTTTTTGCAAATTTTGATGATG
>Dere_Responder
TAAAATCGAAATGATCGTTGGCCACTTACTGACATTTCTGTTCGGCTGGTATCTGAAATTCGAAATATCTTCATT
TTTCGAACAAAACTCGGATTAAAAGTCCGGTTTTTGGCACATTTCGTTGTTTTATGCACCAGGCAATCATAATTA
GCAGTAGGTGGCTAC
>DeugSat1-8
TGTCAGTC
>Deug_1.688
CACATTTTTCGCAATTTTTAATGATGTTACCCCTTACAAAAAATTCAAAATTTTACAAAAATTAAAATCCAAAAA
ATCAACAAACAAATGATGAGGAAAGTTAGTTTTGGTCGTTAGCTTCACAAAACAGTAATTCTTCTAGCTGTACGT
CTTGTTTTGGCTAAGTTATAGGTATGAAACCAATAAAAATTTTGACATTTTCTGTCAAAAGTCTTATCCTGATTT
TTGCCCCGAAAATAATCAGAATTCTTCTCAGAACAAAGCCAATAATAGTCGCAATAAGAAAAGTCATACATCGCT
GAACTCGTAATCAAAAATCCAGAAGAACTGCATCCACACCTTAAAATGTAAAATTTTG
>DeugSat3-112
GGTTTACCATTTTCCTCCTTTAAGTGACAGTATTATGGCTATAACTAAGCCAATAGTAGCCGAATCTGGAATGGC
ATAGTCGTTGGATTCTTTATTAATTTAAACTT
>DficSat1-197
TGTCAAAATTTGCAAAAAATGGGTTTGCAGAAAAGTGACCAGATCCCAGCACTGCTTAGCCCAAACTTTTGAAAT
TTTACCCGATTTAAAAGTGGAATACCTCTCTGAATTTGTTATTAAAATATCTATCTAGCTGCATTATTGGTTTAT
TTTCGAATTTTTGGTCAATATTTTGTAATTTTTTATGACCCCCGACT
>DficSat2-189
GAACAGGGGTCAAAATTTCGACCATCGGAAATGGAAAATTTTGAATGGGGTTGTACCAAGGATGCGATTAGAAAC
GTGTAGAGGTTGCCACTTTGGAAACAGAGTGAAATTGACAAAGTTATGGGTTCTGGAACTCCAGTTTTGGGTCAC
TATATTGAAAACAATCGGACAATTGGAAAAATCTGACTT
>DhydSat1-7
CCTTTGA
>DhydSat2-6
AGAAAG
>DhydSat3-14

AGTTCAAATGAACT
>DhydSat4-2
AC
>DhydSat5-37
CAAAAAGCTGATTGCTATAAATGATACGACTGACTAA
>DkikSat1-41
CTAAATAACGAAGAAGAAGGCAAAAAAGAAGGAGTAGAAGA
>DkikSat2-109
ACTCAGCTAATAATGCATGAATCCACTTCAGAATTTGTATTTATATTAGTTTTTACACGGCTAATACATTTGCAT
ACTCAATTGTAAATTTGTAAACAAAAAGTTCACA
>DkikSat3-19
TGAGCATCACTGATGGAAG
>DleoSat1-41
CCGAAGAAGAAGGCAAAAAAGAAGGAGTAGAAGACTAAATAA
>DleoSat2-19
TCAGTGATGCTCATATGCA
>DleoSat3-111
GGTGGCCCCTAGTCCGAGGACGGCTGCTTCTCCCCGTCAGCAGGCGCTTGGGCGGCTTGGGATGGGAGTGGACGC
GGCAGTGTAGTAGCCACTCGCAGCGGGACCAATGGT
>DleoSat4-109
AATTGTAAATTTGCAAACAAAAAGTTCACAACTCAGCTAATAATGCATGTATCCACTTCAGAATTTGTATCTATA
TTAGTTTTTACATGGCTAATACATTTGCATACTC
>DleoSat5-28
GTATCGGAATAATGTTATCGGAAAACTG
>DleoSat6-570
CTCATGGCACAGATCAGCTCGCTGATTCTTTGAGAGCAACGATTAAAATAACTCAAAGACCCAATGAAAAAGAAC
AACTGTTGTTGTTGTTAAATGCATTTATTGTAGGAATATTGTTTTTATTGTTGTAGAAACAACAAGGCTTCAATT
TGGACATTTTCAGTACAAATTCATACATTTCTGGGAGTCAAGTAACAGGCGAACAGAATCAGCAGGAGGTGGATT
TACATTGGTGCAAATCAATATTTTCACTCAATGTTAATACTTCCACTCATTGTTTTCACTCTTCGCCACAGATCG
GCTCATCTAAGCTCTACCGCTCTGTGTTCGAAAGTTGTTCGAAAATTGTTCGAAACTTGTTCGAAAGTTGTTCGA
AACGGAGAGACGGACCATTTATGAAAGCGGGAGAGCAGCGGTTACCATAACTCTTTGACTCAATGGATTTTCAGA
GCTGCAAGTGACCTACGGTCTCCGCTCTCCTGCTCTTTGCTCGCCAGTTACGGCCTTCGGGGATTTGGAGAGCCA
AAATGAGTGCACAGCATTGTTTTTACTCAAATTCAACATTTCTCA
>DmalSat1-188
ATCCTTCAAAAATCCGAAAAATGGATGTAAGGCCCAATATGGCCCACTGCGAAGGACATATCTTTGTCAATATCC
ATCCGATTCTCAAAAGGAATACCTTAAATGATTTGTGGATCGATTTCGAACAAATCTGCATCAAAACCTGGAAAC
AAATTATTTTTTAGATTTTTTGTCAAATTTTGTGGGGT
>DmalSat2-32
CCTGGCCGATCCCCCTGCCAAGGCGTGGAGTG
>DmalSat3-197
GTACCAGTCCCATTAAAATCGCTTCCACAGATTTTGAGAAATCAAGTCCCGAAATTTTCGGACCCTGATTCGCCA
ATACTGGCCATTATCATTTTCCCGAGTTTTGACGCTGCCCAAATGATTACCTTTTTTTTTACAGAGGATGCGCTAT
AATGCACGGTATACCACGAGAAAGCTAATTTTTAACCCATTCGAATG
>DmalSat4-191
CAGAGCTATGGCCATCGGAAGGAGCAAAAGTGTGGAAATTGTGGAAAAACCACAATGTTGTAAGGGAACTCCTTG
GAAAAATTTCGACAAATTGAAAAACGACATTAATGCAATGTTTTGAATAAGGAAAGTCGCTTTTGGGTTTGCTGA
TTACTAAATGGTACTTTGTTTCCTGATCGGAAACAAATTGG
>DmalSat5-205
AAAATTCGACAGTTCTATTAAAAGAGAACAAAAGACGGGAACTCATCCATTGATTTTTCCAGATTATCTGCTTGT
TCCAAAAGACCCAGAAAAACCATTAAAAAAGAGTCAGCCTAAACAACCTTTTTTAGTGTTTTCGTGTCTCACATT
TATTTGCAAGTTAAGCACAAGAATTTAATGTTATGGAACGTATCGAGACAAAAC
>DmalSat6-380

GAAACACCCAGGAACCACAGCCCGCTGATACCCAGATCCCAGATGCACCGCTGGTATGAGGCAAAGACGGCCAGC
TTTTTCCTACAAACAGAAAACGAAGTTATACCACAGCTCATCTCCAAATTGGCCGACGACTACATACCTGGTTTC
CGGCAACAAATACGCCCAACAAGCACCGGCTTGCCAACTCAACGCCAAATCTCCTGGAAACGCGACTCCTGTCCT
CGCCACCAGGGAAACGTAAGCGCCCAGGCAGCTAGCACACACAATCGGCAACAGCCCCTAGGCCAACAGGAAATC
TGCACCACGGAGGCCAACAGCAAATCCCCCACTCAGGGACACAGCTGCTCAGACGATCTGAAACCAGCACGAAAT
AAGCC
>DmauSat1-15
GAACAGAACATGTTC
>Dmau_ Dodeca
GGGACCGAATAC
>Dmau_1.688
CAATTTTTTGCAAATTTTGATGATGGTACCCCTTACAAAAAATGCGAAAATTTGGCCAAAAATTAATTTTACAAA
ATCCGTTTAAAAGTGAAAGGGGTGGTTAGTATTGGTTATAAGGAGTATAAAATGGTACTTCTTTTTGCTGTCTGA
CCATTTTTAGTCAAGTTATAGCCAAAAAGCCAACTTGAAAATTGAGTTTTTTGCCAAAAATAGTTTTCCAGATTT
TTTGTGAAAAAAATTATCGGTATTTTGATCAAAACATTGAAAATAATGACCCAAATATGGAATGTCATACCTCGT
TGAGTTCGTTGTTTAAATCCCAATCGATTGCATTCAAGTTTGGAAATTCTAGGATTTTT
>DmauSat4-147
AAAAAATATTCTAAAAATATTCGGTAAATATATGTTAAAAATACTAAGTTGACAAACACTAGAAATATATACCAG
AAACAAAAACTATATTTGCACTGTGCCGTGTAGCTTACGTGCCACTAATGAGAATCGAGGAGGAGGGCGCAA
>DmauSat5-197
TCAAATACAAGAAATGTCGTTATATTCAAACTCCTTGTTCGGATTCATAACCACATATTGACATTTGACCAATGT
AAAAATCGTGTATTTACAACTTATGTATATATTCAGCCAATTTGTACAGTAACTTCGGTGAAAATCTTAAGATGA
TAAAACCATCGGTTAAAACGAAATCGTTTTACAGTGGGTGGGGTAAT
>Dmau_Responder
CACAAAAATGCGCATTTTGTAGGCCAAATTGAGAGATTTTTCGGTTTAAACTACCAGGCGAACAAGAATTCCAGA
AGGTGACCAACGATAAGTTGGGTATGACCAACGATAAGTTGGTTTTCAGTAGACACCAGCTACTGATTATCATCG
CCTGGTATCTAAAATCGAAAGATCCAAAAATTTTG
>DmauSat7-39
AACTGAGGGAACAACGGCCAAGCCAACAACTCAAAAACC
>Dmoj_CDSTR130
CAATAGAGGAAAATGAATTCTGTAATTATTGCAACATGTGATGGAGAATCAGAAAAAGCAAATCGATTCAATATT
TTGATGCTTTCATAACAGAAAATTTTAAATATTTGGTAATAAACTAGTAAATATG
>Dmoj_pBuM-1
CCTGAAATGCGGTAAAAGATCAGAAAATTGACTTTTTTCTATGTATAACAGGCAATAACAGGACTATTTCTGGGC
CGATTTCAGTACGACTTACTTCGTTGGAAGCGTCTTTGAGAGCCCCATCCATTGACGTATTCAGATTTTCAATGA
GCCAACAACTTTTTGAAATATTTACATTATTATCCC
> DmonSat1-3
GCA
>DmonSat2-4
AGAC
>DmonSat3-55
GAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG
>Dmon_*pvB370*
TATTTGGCCAAGATATTCCAAAAAATCATCAGAAAAGTTTGACTTTACAAACGAATCGGTTTTTGGTCAACCACT
TTTTTTAAGCCATCGATTTTTAAAATGTTTGAACGCAGATTGATAGTAGGGATCCCTACGAATTCAAAAAGGTAT
AACATTTCAAAATCTGACAATATTTACCCGAGATATTCAACAAAAATCATCAAAAAAGTTTTGATTTTCGAACCG
ACCCCGAATTTGAGTAAAAACGTGATGTAACCCCTTGCCCATTTTGTGCATTTGGCTCAAAATTTAGATTTTTAT
TTTTTTGCATGCCAATCGATAGTACTGATTCTTACGAGTCTAAAATGGTGTAAAATTCCCAAATTGGACAG
>DmonSat5-393

GTTGTATGAAAGTTTTTTGTTTACCATCCGATCGGCCCCAAACTTATTTTCAAAGATCTTTGTTCAAATTCATTT
GAGGGGTGTTTCACACTTTTCGAAAAATCCGGTCCTCCTTGGTGCAAATGGGGGTCGAAGATTTGGTGGATGACG
TTTTGTGTTCAAATTTATTTCCAAAGATCTATACGAAAAATCATTTTTGAGACGTGTTTCACACTTTTCGCAGAA
TTCAAACTTCTTCGGGGGGAATGCGATTTAAAGATTTTTTGGGCGAAGTTTTAAGTTGTGTCCGATTTGCTTCAA
ACTTATTTTCAAAGCTCTACATAAGAACTTAATAAATACGTATTTCAGCGATTTGGGAAATTCCACCCGCAACAG
TGGTAAATTGGCGGAAAACG
>DmonSat6-172
CCCATAACTCGGTCAAATCTCATCCGATTTTCACGAGGTTTGTCTTTTTGTTCATGGTTTTGCCTCTTTATCAAT
TTGGCATCTAAATCTGGCAACATTATTCTTGGTCAAAATTCATGTGAAAATGGTACCCTTAAATATCCTATATAG
ACATAGGTCATAATTTCCCACC
>DnovSat1-172
CCCATAACTCGGTCAAATCTCATCCGATTTCCACGAGGTTTGGCTTTTTATTCATGGTTTTCCCTCTAAAATAAT
CTGGCATTAAAATCTGAACACCTATTTTTTTTGTCTAAATTCATGTCCAAATCTTACCCCGAGATTCGTATATAG
ACATGGGTCAAAAATTGCCACC
>Dnov_*pvB370*
TATTTGGCCGAGATATTCCAAAAAATCATCAAAAAAGGTTGATTTTAAAAACGAATCGTTTTTTGGTCAACAACT
TTTTTTAACCAACCGAATTTTAATTTCTTTGAATGCCAATTGATAGTAGGGATCCGTACGAATTAAAAAAGGTAT
AAAATTACAAAATCAGACGATATTTACCCGAGATATTCAATAAAAATCATCAAAAAAGTTTTGATTTTCGAACCA
ACCCCGAATTTGACAAAAAACGTGATGTAACCCCTTGCTTTTTGGCCGATTTGGGTCAAAATTTGGATTTGCATA
TTCTTAATGCCAATCGATAGTACTGATCCTTACGAGTCCAAAATGGTATAAAATTCCAAAATCTGACAT
>DnovSat3-6
GCGACA
>DnovSat4-190
GTTTTGCCTTGCCGGCATAAGTCAATTATGTTAAACTAAAAAGATTACTGAAACTAATAGAATCTATCACTTAAA
TAGTACATGAAACAAATAAAGTGGTTGGTAGCATTCGGAAAAAATGTCGACATTATAGAATATGTGTCATAAAAC
CAACATAAAAAGTTTCTATATGGCATAACACTGTGTG
>DnovSat5-33
TGCAATAGCTGACAAAAGCTGATTGCTATATG
>DnovSat6-393
TTTGTATTAAACTTATTTGTTTAACATCCGATCGGCCTCAAACTCATTTTCAAAGAAATTTCTTCAAATTAATTT
TAGGGATTTTTCCCATTTTGGAAAATCCATTGCTCTTGGTGGAAATGGGGTCAAGAATTTGATAGGTGACGTTTC
GCGCTGAAATTAATTTTCAAGGATCTATAAGAAAAATCATTTAAGACGTGCTTCACATTTTTCGGAGAATTAAAG
CATCCTCGGTAAAAAGTACGATTCCAACAATTTCTTGTGAAGTTTTAAGCTATGTCCGATTTGCTTCAAACTAAT
TTTCAGAGCTCTACATAAATAAGAACTTAATAAATACGTATATCAGCGATTTGGGAAATTCCACAAACAACAGTG
AGTAATTAGCAAAAAATA
>DnovSat7-100
TTGTGGCTAATTATGCAAAACTAAATGGTTTAGCCTAACCAGGGTTGGAAAAAAGTCAACTTAACCAAAAATCAC
TTGGATTTGCATATAATGTATATTT
>DoreSat1-172
CAATCGATAAAACTATTTTTAAAAGGTAAAAATCCTTTTTTATTGAAAAAAAATAGATTTATGAGGTGGAAAATG
AAAATATGGAATTCAAAATGTTGGGAAACATCAAAAACTATTTGTAATGTCTGGGGAAACGTTAATTATTAAAAA
TAACTAGTTTTTAGAGGTTGG
>Dore_1.688
CCATTCTTTGCAAATTTTGATGATGTTACCCCTTACAAAAAATGCGAAAATTGACCCAAAAACGAATTTCCCAAA
ATCCGTCAAAAAGTGATAGGGATCGTTAGCATTGGTAATTAGCTGCTGAATACAGTTATTATTTCATGATTATGA
TCATTTTTGGCCAAGTTATGATTAAAATGCCCATTTAAAATATTTCAAGTTTTTAAAAAATTCGTTTTTTCCAAA
TTCGGTCATAAAATAATCAATTTTTTTTTCCATAACTTTGAAAATAATTGTCTGAATAAGGAATGTTATATCTCTT
TGAGCTCGTAATTCAATTTCCAATCGAACTGTGTTCAAAAATGGCAATGCTAATTTTTTG
>DoreSat3-181
TTGAATTTCTACAACATCATAACATCAAAAAATTACAAAAAAATTGAAAAAAAATAAATTTCTAAATTTGAATGC
AGATTTTTAGAGAATTGCGCCAGAAACATTTAGAGGTATGCCACTTCAAAATCCGATAAAGATAAGTTAAGTTAT
GGAGCTTAGAACTTCGGCATTGGGACTGATT

>DpseSat1-170
CACCTATATAGCATCCACGCTCCAGGAAATGTGGCAGGATCGCACATCTTTGGCATAAGCACCAATATGGCAATA
TAGCCCCAAAAAAGTGGAACCTCCGCAACTATATGGCACCTGCACTTATATAGAACACTTCCCTATCCAGCACTG
GCATGATACTGAACCACCG
>DpseSat2-21
AACTTATCTCCGCTGGCGGAA
>DpseSat3-310
GGAAAGTCATGCCGCACCCCGATCGGCTGCAGTATGCCAGAGATATAGTATACAGAAGAAACCAGTATTTTACTG
ATGTGCAAATCCAAATCTTCCGAAGGCTATATCTCGGCTAACTATGGGCCGATCGGCGAAGGACCCACTTTCGCA
GGATCGTAGGGATCCAGACTGTCGATGGGCATCAAAATCTCGAATACGAAAATGAGGCCGATTTTTGGCAAATTT
AATGATGTAACCATCATGATTTTTTGCGAAAATCGTGAAAAAATGGATGTCCTTTTTTCCTATGACAGTTGACTG
GCAGGACTCTCCCGATCAC
>DpseSat4-267
TAATGAAAAATTATTAATTTATTTCATATGATATGGGGAATATACTTAGCCGGCAATGTCTGTGTATTTTATATG
AATATATATATATTTCTTAATATATATGAATAATTTTGAAGTGATAGTATTGTATTTTTTATTAAAATACGAAAA
GAAACTATTGTTAAGAGGCACTCTCAAGTAAAAATACCCGATTTGCGGAAAGCGTGGCAGAAAACCTATATAGGG
AGTGGTAGTCTCGTGCATGGTATATTTCATATGAAAATTTTA
>DperSat1-170
CACCTATTTAGCACCCACGCTCCAGGAAATGTGGCAGGATCGCACATCTTGGGCATAAGCACCAATATGGCAATA
TAGCCCCAAAAAAGTGGAACCTCCGCAACTATATGGCACCTGCACTTATATAGAACCCTTCCCTATCCAGCACTG
GCATGTATACTGAACCACCG
>DperSat2-319
GGCATGTCTTTTCGTGACCGGGCGAGTCCTGCCAATCGACTGCCATTGGAAAAAGGGACATCCATATTTTCACGA
TTTTCGCAAAAAATCATGATGGTTACATCATTAAAATTGCCAAAAATCGGCCTCATTTTCGAAGTCGAGATTTTG
ATGCCGTTTGACTCTCTGGATCCCCACGGTCCTGGGAAAGTAGGTCCCACACCGATCTGGCCCTTTTAGTCTGAG
ATATAGCCTACCGAAGATTTGGATTTGCGTATCTTTAAAATACTGGTTTCTTCTGTATGCTATATCTCTGCGATA
CGGCAGCCGATCTGGGCGT
>DperSat3-21
AACTTATCTCCGCTGGCAGAA
>DperSat4-267
TAAGGAAAAATTATTAATTTATTTCATATGATATGGGGAATATACTTAGCCGGCAATGTCTGTGTATTTTATATG
AATATATATATATTTCTTAGTATATATGAATAATTTTGAAGTGATAGTATTGTATTTTTTATTAAAATACGAAAA
GAAACTATTGTTAAGAGGCACTCTCAAGTAAAAATACCCGATTCGCGGAAAGCGTGGCAGAAAACCTATATAGGG
AGTGGTAGTCTCGTGCATGGTATATTTCATATGAAAATTTTA
>DrhoSat1-191
TTTCCGAGGACTTGTAACAGAGTGACTCAAAATTGGACTTAGAGAGTCCATAACTTTACCAAACTTAAGCCGATT
CCAAAGTGGCATACCTCTAAAGACTTCTGATTCGATTCTCTAAAAATCTGCATTCAAATTTTTCAATTTAACGTT
GATCAAATTTTTGACCCATTTTCATGTCAGTTATTTCTAA >DrhoSat2-11TTTCTATAGA
>DsanSat1-9
CACAAGTAT
>Dsan_1.688
ACTTTTCTTCCAATTTTTGATGATGATTTTTTGCCAATAACCGCTTTTCTGTTATTTTGCGAATATAAATATCAG
TATTTTGATCACAACATTCGAAATAGTGGTCCAAATATGGAATGGCATACCTCGTTGAGCTCGTAATTAAATTTC
CTATCGAACTGTGTTTTCAAAAAAAAAAAATACTATTTTC
>DsanSat4-96
ATGATAGAAATAAAGACTACAAAATTCTATCAGGTTGCCAAATAGCTTGTCATCAATTTAGTGACGCATATGAAA
TTGTCCCCTCCCTATCTACTA
>Dsan_Responder

CGATCATTTCGATTTTAGTAGCCACCTACTGCTAATTATGATTGCCTGGTGCATAAAACAATGAAATGTCCCAAA
AACCGGACTTTTAAGCCGAATTTTGGTGGAAAAATGGAGAGATTCTCGATTTCAGATACCAGGCAAACAGAAATG
TCAGGAAGTGGCCAA
>DsanSat7-241
GGCGCCAGTGTTGTTAAAGCTGTTGAGTATAGTGGCGGCAGTGTGGTGTGTTGGTAAAGCAGTTGTGTGTTGTTT
AATACCGCTATCCGCACTGGCGCTACTATACAGAACAGCTTATACAAGTGGTGCCAATAGAAACATCTCAAAAAG
TCGCTCAATCTGCGCTTCTTGTGTAGCAATTCGGTTTCCCGATATGTCGTCTCCTTTCCTTTTCTTTCTTTCCTA
CGCAGTAGAGTGTAGT
>DsecSat1-15
GATCGAACAGAACAT
>Dsec_Responder
CTACTGATTATCATCGCCTGGTATCTAAAATCGAAAGATCCAAAAATTTTGCACAAAAATGCGCATTTTGTAGGC
CAAATTGAGAGATTTTTCGGTTTAAACTACCAGGCGAACAG
>DsecSat3-5
CAGAA
>Dsec_Dodeca
CGGGACCGAGTA
>Dsec_1.688
CAATTTTTTGCAAATTTTGATGATGGTACCCCTTACAAAAAATGCGAAAATTTGGCCAAAAATTAATTTTACAAA
ATCCGTTTAAAAGTGAAAGGGGTGGTTAGTATTGGTTATAAGGAGTATAAAATGGTAATTCTTTTTGCTGTGTGA
CCATTTTTAGTCAAGTTATAGCCAAAAAAGCCAACTTGAAAATTGGGTTTTTTGCCAAAAATAGTTTTCCAGATT
TTTTGTGAAAAAAATTATCGGTATTTTGATCAAAACATTGAAAATAATTACACAAATGCGGAATTTCATACTTCG
CTGAGTTGAGTTTGTTTCTTGAATCCCAATCGAATTGCATTCAAGTTTTGGAATTCTAGGATGTTT
>DsecSat6-193
TTTGTACAGTAACTTCGGAGAAAAGCCTAAGATGATAAAACCATCGGTTAAAACAAAATCGTTTAACAGTGGTTG
TGGGTAATCAAATACAAGAAATGTCGTTAAATTCAAACTCCTTGTTCGGATTCATAACCACATATTGACATTTGA
CCAATGTAAAAATCGTGTATTTACAACTCATATATTCAGTCAA
>DsecSat7-39
CTGAGGGAACTACGGCCAAGCCGACAACTCTGAAACCAA
>Dser_CDSTR198
AATGTAGAAAGGTAGTTGGTGAACTAGCAACAGCTAAATATGGGTAGAAAACAGCCAGAAAGGTATATAACGCAC
TGGTGGGCAATAAATTGGTGGATAAATAAGACTAGCCATCATCCCATAAATAATATCATGATTAAGAGATAGAAA
TATGATTAGAAAGTAGAATAGAAAGGTTAGAAAG
>Dser_CDSTR138
CAACTGTTGACTTTCGTATATTGAATACGAATAAAATCGAGTATGTGGTGTAGAATATTATATTTGCAGCGATTT
TCCATGAACATAACCTTAGTTTGATTTTTTATTTATTTATAGAGAGTATAGGCATATCTGGGC
>Dser_pBuM-2
AACCATCTTAGGTAACGTAAACCCTCTTTACTGTCGCAAAAAGCGGCTTGTGCGAAAAAAAAATTGACTATTTTC
GAGATTTAACAAGCTATAACCGGAGTATTTTTCATTCGACTTCTTTATGGCATACCATTTTAGAAGCGTCCTTCT
ATCTCAGATATATCTAAGGATCCGGATCCTTAGATATATCTAAGGATCTGGCATGGTCTAAGAACTTTCCGAAAG
ATGACCAATTGAACTTTAAATGGCTCTAGCTAGGTTATCCTTTGTCCGATTTTAAGCTCAACTATACTTTTTAAA
TCGACATAGACAGTTCTTTTCATATAAACACAAACATTTTTTTCGAAAAAATTTTCATTTTTGCCATGTT
>Dser_DBC-150
ATATATACCGGGTACTGTCCACGTATTTCCGTTGGGCCGGGTTATAGGTCAAGCCCCAGTCACATATGTATGTAG
CAGGAGGCGGCCGGGTATTGCCGAAATCGCGGTCCGG
>DsimSat1-15
GTTCGAACAGAACAA
>Dsim_Dodeca
CGGGACCGAGTA
>Dsim_Responder
CTACTGATTATCATCGCCTGGTATCTAAAATCGAAAGATCCAAAAATTTTTCACAAAAATGCGCATTTTGTAGGC
CAAATTGAGAGATTTTTCGGTTTAAACTACCAGGCGAACAAG

>Dsim_1.688
CAATTTTTTCCAAATTTTGATGATGGTACCCCTTACAAAAAATTCGAAAATTTGGCCAAAAATTAATTTTACAAA
ATCAGTTTAAAAGTGAAAGGGGTTGTTAGTATTGGTTGTAAGGAGTACAAAATGGTACTCCTTCTTGCTCTCTGA
CCATTTTTAGTCAAGTTATAGCCAAAAAAGCCAATTTGAAATTTCATTTTTTTGCCAAAAATATTTTCCCAGATT
TTTTGTGAAAAAAATATTTGGTTTTTTGATCAAAACATTCGAAATAATTACCCAAATATGGAATGTCATACCTCG
TTGAGTTTGTTTCTTAAATCCCAATCGAATTGCATTCAAGTTTTGAATTCTAGGAGGTTT
>Dsim5-135
AAAAAATATTCGGTAAATATATGTTAAAAATATTAAGTTGACAAACACTAGAAATATATACTAGAAACAAAAACT
ATATTTGCACTGTGCTGCGTAGCTTACGTGCCACTAATGAGAATCGAGGAGGAGGGCGCADsimSat6-
193TCAAATACAAGAAATGTCGTTAAATTGAAACTCCTTGTTCGGATTCATAACCACATATTGACATTTGACCAA
TGTAAAAATCGTGTATTTACAACTTATATATTCAGCCAATTTGTACAGTAACTTCGGTGAAAAGCCTAAGATGAT
AAAACCATCGGTTAAAACAAAATCGTTTTACAGTGGTTGGAGTTAT
>DsimSat7-241
CGCCAGTGGAAACCCAGAGGAACCCACCCACACCACCGGGAAGGATGGCGGCGGGCGTGGAAAGAGGAAAGGGGA
AAGGGGAAAAAGCAAAGAAGAAGGAAGAAAGAGAGAGGAAAGGAGGGGGAGGAGGAGTACGAGGACGCCAAGCTG
ACCCAGTAGCGCATGGACCAGCGAGCGGCGATGGTGGACCCGCCCACACCTCCGAGAAGGAGGACCCCCGGCGGC
TACCCACCCCTCGGCATGGCCAGGAGGAGGTGGTGCCCCAGCCCCAACACTGGCCACAGTCCC
>DsimSat8-39
CTGAGGGAACTACGGCCAAGCCAACAACTTTGAAACCAA
>DsubSat1-12
AGGCAGAGAGAC
>DsubSat2-445
TTCAAGGAACTTTTATCATAATTAATAATTAAAATAAGTTTTTGTACAATATATACGACGAATACTAATATGTAC
ATACATACATATGTAAATACATACACATGTACATACATACATATGTGTATACTTATATATGTACATACATAAATA
TGTAGTACAAACAAATTGTTTTATAATTAATTTGGGTTGCTATTTTACCTAACATAATTGTATGCGTCTGCCTGC
AGCGGCCATTTTGCAGTTGTGTTGTGTTATAGCGCCGGCCATATTTGTTGCCCGAAAAGGGGCGGCGTTTGGGAT
GCAGCGCCCGAGTGTGTATGCAGGCACACGAAAACAACGGCGCTAAGTAAAATAGCAACCCGAATGAATATGAAT
AGCAGCCATCATACAATATCGCAAGTTAGACCCAAATAAATAAAATTACATTCAAATGTTTAAAAAATTA
>DsubSat3-181
TACAAGAAACACAAACACTTCAATATTATTCATATATAGTAACAACCATATACATATCCAACCAAATAAAGTATC
AGACATTGCCAGCTAGGTATTTTCCCCATATCATATGAAAGAAAAATTTAAAATTTTTTAACATATGATATATAC
CAAGCACGACACCACTCCCTATATAGGTTTT
>DsubSat4-106
ATAGTATAGTTTGTGCCAAATTAGACGATCAGGAAAGTTTCAGCTCAAAAGAACACACTTTTGTCAAAAGTCAAA
ATAAGCAGGTATAGACCTTCTAACCCATTTA
>DsubSat5-162
ATATGATATACTGTACTGTATGGCTTTTCATATGAAATAGGCATTGTATTGATATAAATACAAATTTATAATGGT
TGATTGATATGGATATGGGAATATTTCCCATATACATGTATAAAATAACTTCAGCATGATTTGATTATGAAGAAA
CGATAATGTACG
>DsubSat6-226
ATATAAGAAAAATTCAGATAACACACAATAAAGAGGATCTTATATGATATAAGTGGCAATAACTTAGAAAGTTCA
TATCAAATACTTTAAAATACGGGATAAAGAGACTTTGCATATACATATATGCATAAAAACGAATTGAACACTATG
ATACAAATACAAGTCAAAAGTTGATAAACATTTAAAAACACAATAAATGAACATATGTTTCATATATAGTTGGAA
A
>DtakSat1-190
TCTGAAAATAAAGGAAAAACATGGTAACGGATGAAAATGAAAAGACAAGTAGCTAAGACGGCCAATAACTTTTGA
TTGGAATGTCCGATTTAAGCCATAATTGCTGAGTTTATAGGTCTTGATGAGTAGAACAAAGTTGCCATACAATCC
TTCTTTCTATCTCTCGTAGTTTTTTCAAAACCGTCTTAAA
>Dtak_1.688

TTTGAGTCCTCATTTTCCACCCTAAAACATCCGTTTTTTTGCCGAAGCAAAGGGAATAATAGCCCAAATGTGGAA
TGTTATACCTTGTTGAGCTCGTAGCTTAATTCTCAATCGATTTGCGTTCAATTCTGGACATTTTAATTTTTGGTA
ATTTTTTGCAAGAAATCATGATGTAACCCCTTATAAAAAATGCGATAATTAGTGAAAAATCTAATTTTCCAAAAA
GTGATGGGGATCGTTAGTTTAGGTTTGTTAGCTAGTCTAAACAGTCGGGCGCTAAGCTGTCGGCCTTGATTTGTC
AAAATTACGACTGAAAAAGTGTATTTTTTGCCTAAAA
>DtakSat3-33
ATATGGTCTAAGGCGCCCTTTGACCATATTGTA
>DtakSat4-307
CTTAAACAGTCACTTGGATCAAACAAAAAAAGATTTGGGTTGGTGTGCAGGTTAATTAAAAATATTTCGTTTTTC
TCCCAAACGCATAGTTTGGTTATTAAATTAGTGGTGTAAATGCCCTATATTGAAACAAAAACAACGATTTTATTG
TGTTTTGAGCCTAAAACTCTTTCCCATCAGAAAATAATCGTTCAATATGTAAAATTTACAAAACTAGTTTTATTT
TGCTTATCTGAAGAGATATTGTAAATTCAGATGTCTTATTATGATAATCTAATTTAAAATACATAATTTTAGTTG
GTTTTAA
>DtakSat5-132
TTTCCTCTGGAGTGGTTTCGTATTGTTCTGTGCTCGATATTTCAGTGAACTCTGTGATATTGTACTCGATTCGGT
TGAAGACTTTGGTAAGGTTGTACTTGATTGCGGAAGAGATTCCGAAATTGTACTTC
>DteiSat1-10
AATCGAATAT
>DteiSat2-114
AATATGAAAAATGTTTTTATGCTGCAAATTAAGTGTTCAACATGGCATATTAATTTTATGCATATTTATCGTCAT
CTGTTAACAATTCATATTGCGAATTATTTATGAAATATG
>DteiSat3-132
ACATATAACGCACTATTTTAAGATGTACGATAACATTTAATTATTTTATTAATTTCCCGCACCTTAAGTATACAC
CTTGAGAATCGCCGAGAACGGCTGTCCTTCGTAGAAAAAGTATCAGGCAACTTATAC
>Dtei_1.688
TAATTTTTTCCAATTTTTGATGATGATTTTTGGATTTTTTCCGAAAATGGTTTTTCTGTTTTTTTGCGATTTAAA
TATCAGTATTTTGATCAGAACATTCGAAATATGGGTCCAAATATGGAATGTCATATCTCGTTGAATTCGTAATTA
AATTTCCAATCGAACTGTGTTTACAAAAAAAAATGAATTTTTTTTT
>Dtei_Responder
TAAAATCGAAATGATCGTTGGCCACTTACTGACATTTCTGTTCGGCTGGTATCTGAAATTCGAAATATCTTCATT
TTTCGAACAAAACTCGGATTAAAAGTCCGGTTTTTGGCACATTTCGTTGTTTTATGCACCAGGCAATCATAATTA
GCAGTAGGTGGCTAC
>DvirSat1-7
CTACAAA
>Dvir_DNAREP-TR1
AACAAAATATTTTTTCATACAAACTTAATTTTCGACCGATCGTTCCTATGGCAGCTATATGATATAGTGGTCCGA
TCTTAATAGGATTTTGCATATATATGAGGAGTAAAGTAAAACTAATAAATACCGAGTTTGGTCAAGATATCTTGA
AA
>DvirSat2-172
CCCATAACTCGGTCAAATCTCATCCGATTTATACGATGTTTGTCTTTTTATCAATGGTTTTCCCTCTAGAATAAT
CCCTCATTCAAATCTCAACACAAATTTTTTAGCCGAAATTCATGTCAAAATCTTACCCCAAGATTCATATATAGA
CATAGGTCTAAACTTCCCATG
>Dvir_*PvB370*
TATTTGGCCGAGATATTCCAAAAAATCATCAAGAAAGGTTGACTTTACAAACGAATCATTTTCGGATCCACGACT
TTTTGACACCCACCGATTTTCAAGTTGCTTGAATGCCAATTATTAGTAGGGATCCGTACAAATTCAAAAAGGTAT
AATATTTGAAAATCAGACGTTATTTACCCGAGAAACTCAAAAAAAAATCATCAAAAAAGGTTTAATTTTCGAGCC
GACCCCGATTTTGAAAAAAAACGTGATGTAACCCCTTGCCATTTTGGCCGATTTGGGTCAAAATTTAGATTTCCT
ATTTTTGATGCCAATCGATAGAACTGATCCTTACGAGTCAAAAATGGTATAAAATTCCAAAATCGGACAT
>DvirSat4-7
AGACAGA
>DvirSat5-132

GACATAACTCCGCCGCTCTAAAAACGACATATCTCCGCGCTCTAAAAACGACATATATCCGCGCGAAGATATGTC
GTTTCAAAACGACATAACTCCGCCGTTGATCATGGCGAGATTATGTCGTTTTGAAGC
>DvirSat6-397
GTTTGTATTAAATATTTTTGTTAACCATTCGATCGACCTCAAACGCATTTTCAAAGAAATTTGCTTATGTTAATT
TTGGGGTTTTTTCTCCATTTTTAGTAAGTTCCATTGCTCGTTGGTGGAAATGGAGACCAAAAATTTTATAGGTAA
CGTTTCCGCTGAAATAATTTTCAAGCATCTATAAGAAAAATCATTTATGACGTATTTCATTTTTTTCGGAATATT
AAAGCATGCTCGGTAAAGAGTCTGATTCTAAGAATTTTTTGTGAAGTTTTAACTTATGTCCGATTTGCTTCGAAC
TTATTTTCAGAGCTTCACCTTCATAAGAACTTAATAAATAGGTATATCGGCAATTTTGACAATTCCACACACAAC
AGTGGGTAATAAGCAAAAAAT
>Dvir_Tetris-220
ATCATACTGTTAAATGGCAGCCATATTTTAAACTTTATTATTTAATTTCTCCGAAAACGGCCCCAAATGCACGGT
AGTTGTGCATATAGAGCACGATTTATCAATCATTTTGCTATTTATTTCATTAAAATCATCGGTACAGATTTTGAG
AAATTCTCATTTTTCTATTTCTCTCATAACATCCTATGTTAATGCAAAAAAACAAGCTGGCAATATTATAT
>DyakSat1-132
AATTATTTTTAAAGCGTTTAAACCCGTTTTCATTGAATATAGTTAGTTTTAAGAGGTACACAAACTAAGATAATT
ATTTTGAGGGGTGGAAATCGGAAAAAATCGATTTTTAAGGGCAGGCAATCGATACA
>DyakSat2-320
TTGAATTTAATTTCGATTATCAATAAAACGGATGTACTTATTTTAAACATATTATTAATTTCTGAAACTACGAAC
CTGCCTTGTTTTGGTGTAATGTCCAATGCATGACACCAAGAATAAATTAATTCTATCTACAAAATGTAGTTTAGA
TATCAGATGGAAGTTAACCATAAATATTCAACATGATAAATGTTTTTATGCAGCAAATTAAATGTTCAACATAGC
AACATATTAATTTTATGCATATTTATCGGCCATTTGTTTACAATTTATATTGCGAATTATTTGTCTATAATTGTT
TGATTAATATCTAATGTCGA
>DyakSat3-396
CTGCATCAACGTTTATAGAGGCCATTACATTCTCTAGTTTATGGGCAGGTAGAGAGTACATATAAGGAAAAAAAA
CCAGAACAAGTTGGGAATTATTTACAAGCATCATGTATATACTTTTCAAACAATCATTAACAATTACGTGGCCAA
ACTTAACTTGTTTTGACTGTGTATATTTTCGGCAATCAGTTGACATGTCATATTTGGTGTAAGTTGACAGTAATT
CAGTTTTTGTAATATTAAATGTCGATGGGCTTTAATTGTAGATTATTCAATGAACACATATAATTTTAAAAAATA
TAGTTCATTCCACTACTAACCTGGCTTGGTTTGGTCCTGGTTCACGGCGTCCTGAATGCAGATTGTACACCCGCG
CAGCTCCTGAATCCCCATCG
>Dyak_1.688
CCATTTTTCGCAAATTTTGATGATGGTACCCCTTATCGAAAATGCGAAAATTTGTCAAATTTTTTTTTTGCGAAA
ATCGAAAAAGTAGGGATAGACATAGTTAGCTATCTTTATTAGCAGCACAAAACAGTCTTTATTTTAGCTGTGCGT
CCATTTTTAACCAAGTTATGGCCAAAACGCCTATTGAAAATATCCGATTTTTTACAAAAATTTTTTTTTTCGATT
TTTTGATCAAAAATAATCCGTTTTTTTTAAATAACCACAAAAATAGTTGTCCAAAAGTGGAATGCCATACCTCGT
TGAATTCGTAACAAAATTCCCTATCGATCCATGTACATACTTTGAAATTCTAATTGTTTG
>DyakSat5-434
AGGCAGCCAACGATCCCTTTTTTTGCCTAAAATAGAGAGGCTCACGTGTAACTGATTTTTTGTCACGCTTCCAAC
AACCACATGGGATATTTTATTTATGTCGTGATTGTCGTGCAAATCTCGTGAAAAACACGTGAAGCACCAAGGGCA
CAGTGTGACTGCGCATCTCTGATCTGAGTTCCGTTGCAAGTAACGGCATTTCTCTTTAGAGCCTGAAAGCATGCA
AGGGACGGCAATCATGGAAATATCAGATAGTTGTCTCGTTCTATTATAATTGCTGGCATCTGCCGCCACCCGCTG
AGATTTCTGTTCGCCTGGTGTGCGGAACGAAAAACTGGCCGATTTAGCAGCGAAAATAATATTCGGGCGGCTAAA
ATTTTCAAGAAATCGAAGAGATTCCCGATTTTAAGTACCAGGCGAACAAAAATCCCAGT
>DyakSat6-62
GCTGACTTCAGGGTCCTGCTTACGTCTATACGTTGACTTCAGGGTCCTACTTACGTATCTAC

7. CONCLUSÕES

Neste trabalho, combinamos ferramentas de bioinformática, biologia molecular e de citogenética molecular para gerar estudo integrativo e aprofundado dos DNAs satélites em 36 espécies de *Drosophila* com genomas sequenciados. Aqui, descrevemos 172 sequências de DNA satélites, sendo 133 delas descritas pela primeira vez neste trabalho. Também foram realizados experimentos que contribuíram bastante para a caracterização de diferentes famílias de DNAs satélites, analisando os padrões de distribuição cromossômica, conservação entre diferentes espécies, além de um padrão de transcrição diferencial entre os distintos estágios de desenvolvimento das espécies de *Drosophila* cactófilas.

Interessantemente, descrevemos a manutenção das famílias de DNAs satélites compartilhadas entre espécies relacionadas próximas, confirmando a hipótese da biblioteca de DNAs satélites. Além disso, descrevemos a manutenção da família de DNA satélite *1.688* em 13 espécies que compartilharam um ancestral comum há 27 milhões de anos. Até o momento, esta é o maior período de manutenção de uma mesma família DNA satélite no gênero *Drosophila* descrito.

Em relação à influência dos DNAs satélites no processo de evolução do tamanho do genoma de *Drosophila* descrevemos que as alterações no tamanho do genoma do subgênero Sophophora são influenciadas positivamente na variação dos elementos transponíveis, enquanto que nas sequências de DNAs satélites influenciam fortemente a evolução do tamanho do genoma no subgênero de Drosophila. Logo, a evolução do tamanho do genoma dos dois principais subgêneros de Drosophila são diferentemente influencidas pelo parece ocorrer através da modulação de diferentes elementos repetitivos.

De maneira geral, este é o primeiro estudo que aborda a caracterização de DNAs satélites em *Drosophila* de forma abrangente e utilizando um viés de comparação filogenético para a contribuição destes elementos genéticos em um número representativo de espécies. Além disso, neste trabalho foi gerado o maior banco de dados de sequências de DNAs satélites do gênero Drosophila, o qual pode ser utilizado tanto no melhor

entendimento dos DNAs repetitivos deste gênero, como auxiliando enormemente os processos de montagem genêmica.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., ... and George, R. A. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185-2195.

Altemose, N., Miga, K. H., Maggioni, M., & Willard, H. F. (2014). Genomic characterization of large heterochromatic gaps in the human genome assembly. *Plos Computational Biology.* e1003628.

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

Arnason, U., Gretarsdottir, S., & Widegren, B. (1992). Mysticete (baleen whale) relationships based upon the sequence of the common cetacean DNA satellite. *Molecular biology and evolution*, *9*(6), 1018-1028.

Bachmann, L., Raab, M., & Sperlich, D. (1989). Satellite DNA and speciation: a species specific satellite DNA of Drosophila guanchel. *Journal of Zoological Systematics and Evolutionary Research*, *27*(2), 84-93.

Bachmann, L., & Sperlich, D. (1993). Gradual evolution of a specific satellite DNA family in Drosophila ambigua, D. tristis, and D. obscura. *Molecular biology and evolution*, *10*(3), 647-659.

Bachmann, L., Venanzetti, F., & Sbordoni, V. (1996). Tandemly repeated satellite DNA of *Dolichopoda schiavazzii:* A test for models on the evolution of highly repetitive DNA. *Journal of molecular evolution*, *43*(2), 135-144.

Barnes, S. R., Webb, D. A., & Dover, G. (1978). The distribution of satellite and main-band DNA components in the melanogaster species subgroup of Drosophila. *Chromosoma*, *67*(4), 341-363.

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

Bonaccorsi, S., & Lohe, A. (1991). Fine mapping of satellite DNA sequences along the Y chromosome of Drosophila melanogaster: relationships between satellite sequences and fertility factors. *Genetics*, *129*(1), 177-189.

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.

Brutlag, D., Appels, R., Dennis, E. S., & Peacock, W. J. (1977). Highly repeated DNA in Drosophila melanogaster. *Journal of molecular biology*, *112*(1), 31-47.

Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes.

Chen, T., Ueda, Y., Xie, S., & Li, E. (2002). A Novel Dnmt3a Isoform Produced from an Alternative Promoter Localizes to Euchromatin and Its Expression Correlates with Activede Novo Methylation. *Journal of Biological Chemistry*, *277*(41), 38746-38754.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., ... & Civetta, A. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203-218.

Cohen, S., & Segal, D. (2009). Extrachromosomal circular DNA in eukaryotes: possible involvement in the plasticity of tandem repeats.*Cytogenetic and genome research*, *124*(3-4), 327-338.

Craddock, E. M., Gall, J. G., and Jonas, M. 2016 Hawaiian Drosophila genomes: size variation and evolutionary expansions. *Genetica*, *144*(1), 107-124.

Da Lage, J. L., Kergoat, G. J., Maczkowiak, F., Silvain, J. F., Cariou, M. L., & Lachaise, D. (2007). A phylogeny of Drosophilidae using the Amyrel gene: questioning the Drosophila melanogaster species group boundaries. *Journal of Zoological Systematics and Evolutionary Research*, *45*(1), 47-63.

de la Herrán, R., Fontana, F., Lanfredi, M., Congiu, L., Leis, M., Rossi, R., & Garrido-Ramos, M. A. (2001). Slow rates of evolution and sequence homogenization in an ancient satellite DNA family of sturgeons. *Molecular biology and evolution*, *18*(3), 432-436.

de Lima, L. G., Svartman, M., & Kuhn, G. C. (2017). Dissecting the Satellite DNA Landscape in Three Cactophilic Drosophila Sequenced Genomes. *G3: Genes, Genomes, Genetics*, *7*(8), 2831-2843.

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.

Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601-603.

Dover, G. A., & Tautz, D. (1986). Conservation and divergence in multigene families: alternatives to selection and drift. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *312*(1154), 275-289.

Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. *Nature.*229 (5879) 111-117.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792-1797.

Feliciello, I., Akrap, I., & Ugarković, Đ. (2015). Satellite DNA modulates gene expression in the beetle Tribolium castaneum after heat stress. *PLoS genetics*, *11*(8), e1005466.

Ferree, P. M., & Barbash, D. A. (2009). Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS biology*, *7*(10), e1000234.

Gall, J. G., Cohen, E. H., and Polan, M. L. (1971). Repetitive DNA sequences in *Drosophila*. *Chromosoma*, *33*(3), 319-344.

Gall, J. G., and Atherton, D. D. (1974). Satellite DNA sequences in *Drosophila virilis*. *Journal of molecular biology*, *85*(4), 633-664.

Gallach, M. (2014). Recurrent turnover of chromosome-specific satellites in *Drosophila*. *Genome biology and evolution*, *6*(6), 1279-1286.

Garavís, M., Méndez-Lago, M., Gabelica, V., Whitehead, S. L., González, C., and Villasante, A. (2015). The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. *Scientific reports*, *5*.

Garrido-Ramos, M. A. (2015). Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and genome research*, *146*(2), 153-170.

Gilbert, N., Lutz, S., Morrish, T. A., & Moran, J. V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Molecular and cellular biology*, *25*(17), 7780-7795.

Graur, D., Zheng, Y., & Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome biology and evolution*, *7*(3), 642-645.

Gray, D. M., & Skinner, D. M. (1974). A circular dichroism study of the primary structures of three crab satellite DNA's rich in A: T base pairs.*Biopolymers*, *13*(4), 843-852.

Gregory TR. Genome size evolution in animals. *The evolution of the genome* 1 (2005): 4-87.

Gregory, T. R., & Johnston, J. S. (2008). Genome size diversity in the family Drosophilidae. *Heredity*, *101*(3), 228-238.

Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas, F., Puig, M., ... & Ruiz, A. (2015). Genomics of ecological adaptation in cactophilic Drosophila.*Genome biology and evolution*, *7*(1), 349-366.

Hall, S. E., Luo, S., Hall, A. E., & Preuss, D. (2005). Differential rates of local and global homogenization in centromere satellites from Arabidopsis relatives. *Genetics*, *170*(4), 1913-1927.

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, *293*(5532), 1098-1102.

Heslop-Harrison, J. P. (2003). Planning for remodelling: nuclear architecture, chromatin and chromosomes. *Trends in plant science*, *8*(5), 195-197.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome research*, *9*(9), 868-877.

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

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, *110*(1-4), 462-467

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*(2), 111-120.

Khost, D. E., Eickbush, D. G., and Larracuente, A. M. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome research*, *27*(5), 709-721

Kuhn, G. C. S., Bollgönn, S., Sperlich, D., & Bachmann, L. (1999). Characterization of a species-specific satellite DNA of Drosophila buzzatii.*Journal of Zoological Systematics and Evolutionary Research*, *37*(2), 109-112.

Kuhn, G. C., Franco, F. F., Manfrin, M. H., Moreira-Filho, O., & Sene, F. M. (2007). 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 research*, *15*(4), 457-470.

Kuhn, G. C., Sene, F. M., Moreira-Filho, O., Schwarzacher, T., & Heslop-Harrison, J. S. (2008). 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*, *16*(2), 307-324.

Kuhn, G. C. S., Teo, C. H., Schwarzacher, T., & Heslop-Harrison, J. S. (2009). Evolutionary dynamics and sites of illegitimate recombination revealed in the interspersion and sequence junctions of two nonhomologous satellite DNAs in cactophilic Drosophila species. *Heredity*, *102*(5), 453-464.

Kuhn, G. C., Küttler, H., Moreira-Filho, O., & Heslop-Harrison, J. S. (2011). The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes. *Molecular biology and evolution*, msr173.

Laird, C. D., and McCarthy, B. J. (1968). Magnitude of interspecific nucleotide sequence variability in *Drosophila*. *Genetics*, *60*(2), 303.

Larracuente, A. M. (2014). The organization and evolution of the Responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. *BMC evolutionary biology*, *14*(1), 233.

Liebhaber, S. A., Goossens, M., & Kan, Y. W. (1981). Homology and concerted evolution at the alpha 1 and alpha 2 loci of human alpha-globin.*Nature*, *290*(5801), 26-29.

Lohe, A. L. L. A. N., & Roberts, P. A. U. L. (1988). Evolution of satellite DNA sequences in Drosophila. *Heterochromatin: Molecular and Structural Aspects*, 148-186.

Lohe, A. R., Hilliker, A. J., & Roberts, P. A. (1993). Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Genetics*,*134*(4), 1149.

Masumoto, H., Masukata, H., Muro, Y., Nozaki, N., & Okazaki, T. (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *The Journal of Cell Biology*,*109*(5), 1963-1973.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., & Chan, S. W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution.*Genome Biol*, *14*(1), R10.

Menon, D. U., Coarfa, C., Xiao, W., Gunaratne, P. H., & Meller, V. H. (2014). siRNAs from an X-linked satellite repeat promote X-chromosome recognition in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *111*(46), 16460-16465.

Morgante, M., Jurman, I., Shi, L., Zhu, T., Keim, P., & Rafalski, J. A. (1997). The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity. *Chromosome Research*, *5*(6), 363-373.

Mravinac, B., Plohl, M., & Ugarković, Đ. (2004). Conserved patterns in the evolution of Tribolium satellite DNAs. *Gene*, *332*, 169-177.

Nagaki, K., Tsujimoto, H., & Sasakuma, T. (1999). A novel repetitive sequence, termed the JNK repeat family, located on an extra heterochromatic region of chromosome 2R of Japanese rye. *Chromosome Research*, *7*(2), 95-102.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.

Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, *11*(1), 378.

Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, *29*(6), 792-793.

O'Grady, P. M., & Markow, T. A. (2009). Phylogenetic taxonomy in Drosophila: problems and prospects. *Fly*, *3*(1), 10-14.Powell & DeSalle,1995; Powell 1997

Orgel LE, Crick FH (1980). Selfish DNA: the ultimate parasite. Nature. 284(5757): 604-607.

Pezer, Ž., Brajković, J., Feliciello, I., & Ugarković, Đ. (2011). Transcription of satellite DNAs in insects. In *Long Non-Coding RNAs* (pp. 161-178). Springer Berlin Heidelberg.

Pezer, Ž., Brajković, J., Feliciello, I., & Ugarković, Đ. (2012). Satellite DNA-mediated effects on genome regulation.

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., and Mravinac, B. (2012). Satellite DNA evolution. In *Repetitive DNA* (Vol. 7, pp. 126-152). Karger Publishers.

Pons, J., Bruvo, B., Petitpierre, E., Plohl, M., Ugarkovic, D., & Juan, C. (2004). Complex structural features of satellite DNA sequences in the genus Pimelia (Coleoptera: Tenebrionidae): random differential amplification from a common 'satellite DNA library'. *Heredity*, *92*(5), 418-427.

Powell, J. R. (1997). Progress and prospects in evolutionary biology: the Drosophila model (Oxford series in ecology & evolution).

Powell, J. R., & DeSalle, R. (1995). Drosophila molecular phylogenies and their uses. *Evolutionary biology*, *28*, 87-87.

Renault, S., Rouleux-Bonnin, F., Periquet, G., & Bigot, Y. (1999). Satellite DNA transcription in Diadromus pulchellus (Hymenoptera). *Insect biochemistry and molecular biology*, *29*(2), 103-111.

Renkawitz, R. (1979). Isolation of twelve satellite DNAs from *Drosophila* hydei. *International Journal of Biological Macromolecules*, *1*(3), 133-136.

Richard, G. F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686-727.

Rošić, S., Köhler, F., & Erhardt, S. (2014). Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. *The Journal of cell biology*, *207*(3), 335-349.

Rudd, M. K., Wray, G. A., & Willard, H. F. (2006). The evolutionary dynamics of α-satellite. *Genome research*, *16*(1), 88-96.

Russo, C. A., Mello, B., Frazão, A., & Voloch, C. M. (2013). Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zoological Journal of the Linnean Society*, *169*(4), 765-775.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425.

Schwarzacher, T., & Heslop-Harrison, P. (2000). *Practical in situ hybridization*. BIOS Scientific Publishers Ltd.

Scott AF, Heath P, Trusko S, Boyer SH, Prass W, Goodman M, Czelusniak J, Chang LYE, Slightom JL (1984) The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. Mol Biol Evol 1:371-389

Shiels, C., Coutelle, C., & Huxley, C. (1997). Contiguous arrays of satellites 1, 3, and β form a 1.5-Mb domain on chromosome 22p. *Genomics*, *44*(1), 35-44.

Slamovits, C. H., Cook, J. A., Lessa, E. P., & Rossi, M. S. (2001). Recurrent amplifications and deletions of satellite DNA accompanied chromosomal diversification in South American tuco-tucos (genus Ctenomys, Rodentia: Octodontidae): a phylogenetic approach. *Molecular Biology and Evolution*,*18*(9), 1708-1719.

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science*, *191*(4227), 528-535.

Sun, F. L., Cuaycong, M. H., Craig, C. A., Wallrath, L. L., Locke, J., & Elgin, S. C. (2000). The fourth chromosome of Drosophila melanogaster: interspersed euchromatic and heterochromatic domains. *Proceedings of the National Academy of*

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution*, *33*(7), 1870-1874.

Tyler-Smith, C., & Willard, H. F. (1993). Mammalian chromosome structure.*Current opinion in genetics & development*, *3*(3), 390-397.

Ugarković, Đ., & Plohl, M. (2002). Variation in satellite DNA profiles—causes and effects. *The EMBO journal*, *21*(22), 5955-5959.

Ugarkovic, D. (2005). Functional elements residing within satellite DNAs.*EMBO reports*, *6*(11), 1035-1039.

Val, F. C., Vilela, C. R., & Marques, M. D. (1981). Drosophilidae of the Neotropical region. *The genetics and biology of Drosophila*, *3*, 123-168.

Van Der Linde, K., Houle, D., Spicer, G. S., & Steppan, S. J. (2010). A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genetics research*, *92*(1), 25-38.

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I., & Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *science*, *297*(5588), 1833-1837.

Warburton, P. E., Hasson, D., Guillem, F., Lescale, C., Jin, X., & Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. *BMC genomics*, *9*(1), 533.

Waring, M., & Britten, R. J. (1966). Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science*, *154*(3750), 791-794.

Waring, G. L., & Pollack, J. C. (1987). Cloning and characterization of a dispersed, multicopy, X chromosome sequence in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *84*(9), 2843-2847.

Willard, H. F., & Waye, J. S. (1987). Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in Genetics*, *3*, 192-198.

Zhang, H., Koblížková, A., Wang, K., Gong, Z., Oliveira, L., Torres, G. A., ... & Jiang, J. (2014). Boom-bust turnovers of megabase-sized centromeric DNA in Solanum species: rapid evolution of DNA sequences associated with centromeres. *The Plant Cell*, *26*(4), 1436-1447.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., & Wilson, A. C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences*, *77*(4), 2158-2162.

9. ANEXOS

9.1. Pipeline contendo a ordem dos scripts em BioPerl utilizados para a identificação inicial das sequências altamente repetitivas presentes nos contigs montados gerados com a plataforma 454 no genoma de *Drosophila buzzatii*.

==>1
java -jar ~/bin/readseq.jar -inform fasta -f fasta -o Allcontigsteste1.fnt 454AllContigs.fna

../trf_wrapper-01.pl -file Allcontigsteste1.fnt

output:
Allcontigsteste1.fnt.1.1.2.80.5.200.750.dat
Allcontigsteste1.fnt.1.1.2.80.5.200.750.summary.html

==>2 (adotamos o corte de 50pb)
../trf-parser-and-consensus_generation-05.pl Allcontigsteste1.fnt.1.1.2.80.5.200.750.dat repeats-50bp_cut.fasta 50
java -jar ~/bin/readseq.jar -inform fasta -f fasta -o repeats-50bp_cut.fasta2 repeats-50bp_cut.fasta

==>3 (considerando que os arquivos foram gerados no diretorio de trabalho)

/home/trepeats/bin/split-fasta.pl repeats-50bp_cut.fasta2 .

==>4 (deve ser executado no diretorio onde o split-fasta.pl gerou os arquivos individuais)

gerando arquivo de entrada para o blast:
grep '>' repeats-50bp_cut.fasta2 >repeats-50bp_cut_nosign
perl -pi -e 's/ length.*//g' repeats-50bp_cut_nosign
perl -pi -e 's/\>//g' repeats-50bp_cut_nosign

Usage: ../combinatorial-4-bl2seq-02.pl <s><in file>
../combinatorial-4-bl2seq-02.pl 2 repeats-50bp_cut_nosign

arquivo de saida gerado do blast = all_repeats.bl2seq-blastn_out.FW7v40b40 (28Gb)

==>5 (parseamento do resultado blast)
Usage: ../blast_parser_4_clustering_coverage.pl all_repeats.bl2seq-blastn_out.FW7v40b40 all_repeats.bl2seq-blastn_out.FW7v40b40.out.tab 50

==>6 (removendo redundancia na localizacao dos alinhamentos resultantes do bl2seq)
./groups-4-cap3.pl all_repeats.bl2seq-blastn_out.FW7v40b40.out.tab

diretorios 4-clustering; uniq_sequences; temp_files sao criados com os arquivos intermediarios e de resultado da execucao

os resultados estao em: /home/trepeats/data/bl2seq

==>7 (2nd round de clusterizacao)
os resultados obtidos da etapa anterior foram clusterizados novamente no diretorio:

/home/trepeats/data/bl2seq/2nd-round-cluster

relatorio: (file: /home/trepeats/data/bl2seq/2nd-round-cluster/report.txt

ALL-input_several-4-clustering.renamed1:3767
ALL-singlets-4-clustering.renamed2:7518
ALL-singlets-input_sequences.renamed2:614
ALL-uniq-sequences-input_uniq.renamed1:102
TOTAL:12001

after second round
grep -c '>' ALL-ALL.fasta.cap.contigs
782
grep -c '>' ALL-ALL.fasta.cap.singlets
700
TOTAL:1482

Um arquivo FINAL.fasta foi gerado com essas sequencias localizado em:
/home/trepeats/data/bl2seq/2nd-round-cluster

==>8 (blast do arquivo FINAL.fasta contra o genoma de interesse D.buzzatii)

resulatdo em:
/home/trepeats/data/Blastfinal

blastall -p blastn -d Blast-DB/454AllContigs -i FINAL.fasta -o FINAL--vs--454Allcontigs_blastn_1e-5_W7_FF -e 0.00001 -F F -W 7 -a 20

==>9 (tratando o resultado do blast gerado na etapa acima)

./gera-tabela-final.pl FINAL--vs--454Allcontigs_blastn_1e-5_W7_FF

gera um arquivo com o nome:TEMPORARIO-1-all.TXT, veja o diretorio:

/home/trepeats/data/Blastfinal/parserblastfinal

cut -f1,5,7,8 TEMPORARIO-1-all.TXT >colunas-1-5-7-8.txt #query subject_name coordI coordF

cut -f 1 colunas-1-5-7-8.txt | sort -n | uniq -c >coluna-1-uniq.txt #para gera um arquivo contendo o nome de todas as queries

perl -pi -e 's/^[\s]+[0-9]+ //g' coluna-1-uniq.txt #remove a contagem inicial gerada pelo comando sort uniq

./gera-grep.pl coluna-1-uniq.txt saida >roda-grep.bsh
#pega os nomes das queries, realiza um grep no arquivo colunas-1-5-7-8.txt gerando arquivos com todas as linhas em que uma dada query aparece
veja:

/home/trepeats/data/Blastfinal/parserblastfinal/do-jeronimo/tudo-novo/resultado-grep

chmod +x roda-grep.bsh

./roda-grep.bsh

==>10 (analise final de resultados)
veja:/home/trepeats/data/Blastfinal/parserblastfinal/do-jeronimo/tudo-novo/resultado-grep
for i in `ls *-lines.txt`; do ./parser-4-cut_result.pl $i >>repeats-Dbuzzatii.txt; done

cut -d' ' -f 1 repeats-Dbuzzatii.txt | sort -n | uniq -c >cut -d' ' -f 1 repeats-Dbuzzatii.txt | sort -n | uniq -c >repeats-Dbuzzatii--contando.txt
perl -pi -e 's/^[\s]+//g' repeats-Dbuzzatii--contando.txt

sort -k 1 -n repeats-Dbuzzatii--contando.txt >repeats-Dbuzzatii--contando--sorted.txt

executamos o script gccal.pl para calcular o conteudo GC

9.2. Artigo publicado no qual dados obtidos em *D. buzzatii* foram utilizados como parte integrante dos resultados.

GBE

Genomics of Ecological Adaptation in Cactophilic *Drosophila*

Yolanda Guillén[1], Núria Rius[1], Alejandra Delprat[1], Anna Williford[2], Francesc Muyas[1], Marta Puig[1], Sònia Casillas[1,3], Miquel Ràmia[1,3], Raquel Egea[1,3], Barbara Negre[4,5], Gisela Mir[6,7], Jordi Camps[8], Valentí Moncunill[9], Francisco J. Ruiz-Ruano[10], Josefa Cabrero[10], Leonardo G. de Lima[11], Guilherme B. Dias[11], Jeronimo C. Ruiz[12], Aurélie Kapusta[13], Jordi Garcia-Mas[6], Marta Gut[8], Ivo G. Gut[8], David Torrents[9], Juan P. Camacho[10], Gustavo C.S. Kuhn[11], Cédric Feschotte[13], Andrew G. Clark[14], Esther Betrán[2], Antonio Barbadilla[1,3], and Alfredo Ruiz[1,*]

[1]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Spain

[2]Department of Biology, University of Texas at Arlington

[3]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Spain

[4]EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), Barcelona, Spain

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[6]IRTA, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Barcelona, Spain

[7]The Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia

[8]Centro Nacional de Análisis Genómico (CNAG), Parc Cientific de Barcelona, Torre I, Barcelona, Spain

[9]Barcelona Supercomputing Center (BSC), Edifici TG (Torre Girona), Barcelona, Spain and Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[10]Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Spain

[11]Instituto de Ciências Biológicas, Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[12]Informática de Biossistemas, Centro de Pesquisas René Rachou—Fiocruz Minas, Belo Horizonte, MG, Brazil

[13]Department of Human Genetics, University of Utah School of Medicine

[14]Department of Molecular Biology and Genetics, Cornell University

*Corresponding author: E-mail: alfredo.ruiz@uab.cat.

Accepted: December 23, 2014

Abstract

Cactophilic *Drosophila* species provide a valuable model to study gene–environment interactions and ecological adaptation. *Drosophila buzzatii* and *Drosophila mojavensis* are two cactophilic species that belong to the *repleta* group, but have very different geographical distributions and primary host plants. To investigate the genomic basis of ecological adaptation, we sequenced the genome and developmental transcriptome of *D. buzzatii* and compared its gene content with that of *D. mojavensis* and two other noncactophilic *Drosophila* species in the same subgenus. The newly sequenced *D. buzzatii* genome (161.5 Mb) comprises 826 scaffolds (>3 kb) and contains 13,657 annotated protein-coding genes. Using RNA sequencing data of five life-stages we found expression of 15,026 genes, 80% protein-coding genes, and 20% noncoding RNA genes. In total, we detected 1,294 genes putatively under positive selection. Interestingly, among genes under positive selection in the *D. mojavensis* lineage, there is an excess of genes involved in metabolism of heterocyclic compounds that are abundant in *Stenocereus cacti* and toxic to nonresident *Drosophila* species. We found 117 orphan genes in the shared *D. buzzatii*–*D. mojavensis* lineage. In addition, gene duplication analysis identified lineage-specific expanded families with functional annotations associated with proteolysis, zinc ion binding, chitin binding, sensory perception, ethanol tolerance, immunity, physiology, and reproduction. In summary, we identified genetic signatures of adaptation in the shared *D. buzzatii*–*D. mojavensis* lineage, and in the two separate *D. buzzatii* and *D. mojavensis* lineages. Many of the novel lineage-specific genomic features are promising candidates for explaining the adaptation of these species to their distinct ecological niches.

Key words: cactophilic *Drosophila*, genome sequence, ecological adaptation, positive selection, orphan genes, gene duplication.

9.3. Artigo como primeiro autor publicado na revista DNA Research.

Dna Research, 2017, 0(0), 1–9
doi: 10.1093/dnares/dsx010
Full Paper

OXFORD

Full Paper

Identification and characterization of a subtelomeric satellite DNA in Callitrichini monkeys

Naiara Pereira Araújo[1,†], Leonardo Gomes de Lima[1,†], Guilherme Borges Dias[1], Gustavo Campos Silva Kuhn[1], Alan Lane de Melo[2], Yatiyo Yonenaga-Yassuda[3], Roscoe Stanyon[4], and Marta Svartman[1,*]

[1]Universidade Federal de Minas Gerais, Laboratório de Citogenômica Evolutiva, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Avenida Presidente Antônio Carlos, 6627 - Pampulha, 31270-901, Belo Horizonte, Brazil, [2]Universidade Federal de Minas Gerais, Laboratório de Taxonomia e Biologia de Invertebrados, Departamento de Parasitologia, Instituto de Ciências Biológicas, Belo Horizonte, Brazil, [3]Universidade de São Paulo, Laboratório de Citogenética de Vertebrados, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, São Paulo, Brazil, and [4]University of Florence, Department of Biology, Florence, Italy

*To whom correspondence should be addressed. Tel. +5531 34092965. Email: svartmanm@icb.ufmg.br

†These authors contributed equally to this work.

Edited by Dr. Minoru Yoshida

Received 16 September 2016; Editorial decision 27 January 2017; Accepted 15 February 2017

Abstract

Repetitive DNAs are abundant fast-evolving components of eukaryotic genomes, which often possess important structural and functional roles. Despite their ubiquity, repetitive DNAs are poorly studied when compared with the genic fraction of genomes. Here, we took advantage of the availability of the sequenced genome of the common marmoset *Callithrix jacchus* to assess its satellite DNAs (satDNAs) and their distribution in Callitrichini. After clustering analysis of all reads and comparisons by similarity, we identified a satDNA composed by 171 bp motifs, named MarmoSAT, which composes 1.09% of the *C. jacchus* genome. Fluorescent *in situ* hybridization on chromosomes of species from the genera *Callithrix*, *Mico* and *Callimico* showed that MarmoSAT had a subtelomeric location. In addition to the common monomeric, we found that MarmoSAT was also organized in higher-order repeats of 338 bp in *Callimico goeldii*. Our phylogenetic analyses showed that MarmoSAT repeats from *C. jacchus* lack chromosome-specific features, suggesting exchange events among subterminal regions of non-homologous chromosomes. MarmoSAT is transcribed in several tissues of *C. jacchus*, with the highest transcription levels in spleen, thymus and heart. The transcription profile and subtelomeric location suggest that MarmoSAT may be involved in the regulation of telomerase and modulation of telomeric chromatin.

Key words: heterochromatin, repetitive DNA, Platyrrhini

9.4. Artigo publicado na revista Scientific Reports como co-autor.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*

Received: 14 February 2017
Accepted: 19 June 2017
Published online: 25 July 2017

Octavio Manuel Palacios-Gimenez[1], Guilherme Borges Dias[2], Leonardo Gomes de Lima[2], Gustavo Campos e Silva Kuhn[2], Érica Ramos[3], Cesar Martins[3] & Diogo Cavalcanti Cabral-de-Mello[1]

Satellite DNAs (satDNAs) constitute large portion of eukaryote genomes, comprising non-protein-coding sequences tandemly repeated. They are mostly found in heterochromatic regions of chromosomes such as around centromere or near telomeres, in intercalary heterochromatin, and often in non-recombining segments of sex chromosomes. We examined the satellitome in the cricket *Eneoptera surinamensis* (2n = 9, neo-X_1X_2Y, males) to characterize the molecular evolution of its neo-sex chromosomes. To achieve this, we analyzed illumina reads using graph-based clustering and complementary analyses. We found an unusually high number of 45 families of satDNAs, ranging from 4 bp to 517 bp, accounting for about 14% of the genome and showing different modular structures and high diversity of arrays. FISH mapping revealed that satDNAs are located mostly in C-positive pericentromeric regions of the chromosomes. SatDNAs enrichment was also observed in the neo-sex chromosomes in comparison to autosomes. Especially astonishing accumulation of satDNAs loci was found in the highly differentiated neo-Y, including 39 satDNAs over-represented in this chromosome, which is the greatest satDNAs diversity yet reported for sex chromosomes. Our results suggest possible involvement of satDNAs in genome increasing and in molecular differentiation of the neo-sex chromosomes in this species, contributing to the understanding of sex chromosome composition and evolution in Orthoptera.

Among the repetitive sequences in the genomes of eukaryotes, tandem repeats (TRs) are very abundant and are mostly represented by satellite DNAs (satDNAs). SatDNA sequences are mainly located in centromeric, telomeric or intercalary heterochromatin[1–3] but, in some cases, also dispersed in euchromatin[4,5]. This genomic fraction is composed of hundreds to thousands of noncoding tandemly-arrayed sequences with late-replication, and oriented in a head-to-tail fashion[1,3,6–9].

SatDNA families, in general, differ in sequence identity, copy number and chromosome distribution[2,10–12]. These sequences are subject to intragenomic concerted evolution, resulting in more efficient homogenization of repeats within species than between species and also between repeats located in the same array/chromosome than between different ones[2–4,7]. Concerted evolution is achieved through multiple mechanisms of non-reciprocal transfer such as unequal cross-over, gene conversion, rolling-circle replication and transposition[13,14].

Sex chromosomes have arisen independently several times in a wide range of animals and plants from an ordinary autosomal pair[15,16], presenting as a recurrent trait the suppression of recombination and accumulation of distinct classes of repetitive DNAs, including satDNAs[1,17–21]. In Orthoptera, the X0♂/XX♀ sex-determining

9.5. ANEXO 5. Artigo publicado na revista BMC Genomics como co-autor.

Soares et al. BMC Genomics (2015) 16:376
DOI 10.1186/s12864-015-1564-7

BMC Genomics

RESEARCH ARTICLE Open Access

Identification and characterization of expressed retrotransposons in the genome of the *Paracoccidioides* species complex

Marco Aurélio Soares[1†], Roberta Amália de Carvalho Araújo[1†], Marjorie Mendes Marini[3], Luciana Márcia de Oliveira[2,6], Leonardo Gomes de Lima[7], Viviane de Souza Alves[1], Maria Sueli Soares Felipe[4], Marcelo Macedo Brigido[4], Celia Maria de Almeida Soares[5], Jose Franco da Silveira[3], Jeronimo Conceição Ruiz[6] and Patrícia Silva Cisalpino[1,2*]

Abstract

Background: Species from the *Paracoccidioides* complex are thermally dimorphic fungi and the causative agents of paracoccidioidomycosis, a deep fungal infection that is the most prevalent systemic mycosis in Latin America and represents the most important cause of death in immunocompetent individuals with systemic mycosis in Brazil. We previously described the identification of eight new families of DNA transposons in *Paracoccidioides* genomes. In this work, we aimed to identify potentially active retrotransposons in *Paracoccidioides* genomes.

Results: We identified five different retrotransposon families (four LTR-like and one LINE-like element) in the genomes of three *Paracoccidioides* isolates. Retrotransposons were present in all of the genomes analyzed. *P. brasiliensis* and *P. lutzii* species harbored the same retrotransposon lineages but differed in their copy numbers. In the Pb01, Pb03 and Pb18 genomes, the number of LTR retrotransposons was higher than the number of LINE-like elements, and the LINE-like element RtPc5 was transcribed in *Paracoccidioides lutzii* (Pb01) but could not be detected in *P. brasiliensis* (Pb03 and Pb18) by semi-quantitative RT-PCR.

Conclusion: Five new potentially active retrotransposons have been identified in the genomic assemblies of the *Paracoccidioides* species complex using a combined computational and experimental approach. The distribution across the two known species, *P. brasiliensis* and *P. lutzii*, and phylogenetics analysis indicate that these elements could have been acquired before speciation occurred. The presence of active retrotransposons in the genome may have implications regarding the evolution and genetic diversification of the *Paracoccidioides* genus.

Background

Transposable elements (TEs) have been found in virtually all eukaryotic species investigated to date [1,2] and may represent a significant portion of the genomes of living organisms. TEs can account for 80% or more of total genomic DNA in plants and comprise 45% and 20% of the genomes of metazoans and fungi, respectively [3,4]. TEs are DNA sequences with the ability to move from one genomic location to another and can be grouped in two classes according to whether their transposition intermediate is RNA (class I or retrotransposons) or DNA (class II or DNA transposons) [2,5].

Retrotransposons replicate by a "copy and paste" process, whereby the RNA intermediate is reverse-transcribed into double-stranded (ds) DNA by enzymes encoded by the TEs themselves. Elements belonging to class I are further divided into five orders based on their mechanistic features, organization and reverse transcriptase phylogeny: LTR retrotransposons, DIRS-like elements, Penelope-like elements, LINEs and SINEs [2,6]. LTR retrotransposons are the most widespread, especially those from the *Gypsy* and *Copia* superfamilies. Members of the LTR order usually encode

* Correspondence: pscisalpg@icb.ufmg.br
†Equal contributors
[1]Departamento de Microbiologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil
[2]Programa de Pós-graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil
Full list of author information is available at the end of the article