G3: Genes|Genomes|Genetics Early Online, published on June 28, 2017 as doi:10.1534/g3.117.042093

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.

Running Title: Satellite DNA in Cactophilic Drosophila

Corresponding author: Leonardo G. de Lima

E-mail: leonardogdlima@gmail.com

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.

Telephone: +5531-34092612/+5531-992549563

FAX: +5531-34092567

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* reads (76x coverage) were generated by Prof. Alfredo Ruiz group at Universitat Autonoma de Barcelona and were used for the genome assembly of *D. buzzatii* (Guillen et al. 2015). All *D. buzzatii* Illumina reads used on this paper were downloaded directly from the *Drosophila buzzatii* genome project webpage (http://dbuz.uab.cat). This data is publically available for download on the FTP section (http://dbuz.uab.cat/ftp.php); Moreover, we used *D. mojavensis* (SRX2932915) sequence reads (20x coverage) 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

202 characteristic of satDNAs families (Novák et al. 2013). The *Dotlet* software (Junier and
203 Pagni 2000) was also used to generate a scrutinized description of full length copies of
204 each satDNA family.
205
206 **Sequence and phylogenetic analysis**
207 Multiple satDNAs sequences were aligned with the *Muscle* algorithm (Edgar
208 2004) of the MEGA5.05 software (Tamura et al. 2011), with manually optimization
209 when necessary. MEGA5.05 was also used for the analysis of nucleotide composition
210 and variability. Phylogenetic trees were constructed with the Neighbor Joining
211 algorithm (Saitou and Nei 1987) of the MEGA program 5:05 (Tamura et al. 2011). The
212 genetic distance between sequences was calculated using the "Tamura-Nei model"
213 (Tamura and Nei 1993) after an analysis of best substitution model for the data on
214 MEGA 5.05 (Tamura et al. 2011). Statistical evaluation of each branch of the tree was
215 performed using analysis "bootstrap" (1,000 replicates).
216
217 **Samples, DNA extractions, PCR amplifications, cloning and sequencing**
218 For our experimental data we used DNA from the same sequenced strains: *D.*
219 *buzzatii* (strain: ST01), *D. seriema* (strain: D73C3B) and *D. mojavensis* (strain: CI 12
220 IB -4 g8). DNA extraction of 30-50 adult flies was performed with the Wizard®
221 Genomic DNA Purification kit (Promega). PCR reactions consisted of an initial
222 denaturation step of 94 ºC for 3 min, followed by 30 cycles of 94 ºC for 60 sec, 55 ºC
223 for 60 sec and 72 ºC for 60 sec and then a final extension at 72 ºC for 10 min. The
224 primers used for satDNA amplification are listed on Table S1. PCR products were
225 excised from 1% agarose gels and purified with the Wizard SV Gel and PCR Clean-up
226 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.

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

435 Based on the collection and chromosome distribution of the satDNAs
436 discussed herein, the centromeric regions of the X chromosome of *D. buzzatii*, of the X
437 and Y of *D. seriema* or of the Y of *D. mojavensis* are not composed of satDNAs. Some
438 centromeres described in plants and animals are composed of transposable elements
439 (reviewed by Plohl et al. 2014). In *Drosophila*, DINE-1 elements (helitrons) are one of
440 the most abundant types of transposable elements (Yang and Barbash 2008). Kuhn and
441 Heslop-Harrison (2011) and Dias et al. (2015) showed by FISH on mitotic
442 chromosomes that these elements are highly enriched in the sex chromosomes
443 (including the centromeric regions) in the three analyzed species from the *repleta* and
444 *virilis* groups. It is possible that these DINE-1 elements are the main components of the
445 centromeres of the sex chromosomes of cactophilic *Drosophila* species.

446 According to *RepeatExplorer*, the genomic proportion of satDNA in *D.*
447 *mojavensis* (*CDSTR130*+*pBuM*) is 2.5% (Table 1). This value is very close to the 2%
448 satDNA contribution estimated by Bosco et al. (2007) using flow cytometry in the same
449 species. According to the authors, if we split the ~2% satDNA evenly among the *D.*
450 *mojavensis* chromosomes that would result in ~430 kb for each centromere. As noted by

451 the authors, this value is also very close to what is considered as the minimum amount
452 of centromeric DNA (420kb) needed to fulfill centromeric function in *Drosophila* (Sun
453 et al. 1993). In this context, Bosco et al (2007) emphasized that it would be valuable to
454 identify the centromeric satDNA of *D. mojavensis* and other *Drosophila* species to
455 investigate whether they agree with the ~420kb limit observed in *D. melanogaster*.

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

466

467 **New insigths on *pBuM* distribution and evolution**

468 According to previous data on the distribution of *pBuM*-1 *alpha* and *pBuM*-2
469 *alpha/beta* repeats in the phylogeny of *Drosophila* species from the *buzzatii* cluster
470 (*repleta* group), it was proposed that the ancestral state of the *pBuM* satDNA family
471 consisted of *alpha* tandem repetition units around 190bp long. The *alpha/beta* repeats
472 would have been originated subsequently from an insertion of a non-homologous
473 sequence of 180 bp (*beta*) in an *alpha* array, resulting in a composite *alpha/beta* repeat
474 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

549 The *CDSTR198* satDNA was found in *D. buzzatii* and *D. seriema*, but with
550 marked quantitative differences (0.23% in *D. buzzatii* and 0.02% in *D. seriema*). FISH
551 on *D. buzzatii* mitotic chromosomes revealed that this satDNA is located in the terminal
552 regions of chromosomes 2, 3, 4, 5 and X but also spread along euchromatic regions
553 (Figure 4a). FISH on polytene chromosomes of the same species revealed strong
554 hybridization signals in the telomeric regions of chromosomes 2, 5 and X, and in
555 subtelomeric regions of chromosomes 3 and 4 (Figure 3a). Moreover, we detected the
556 presence of *CDSTR198* repeats along euchromatic regions of all chromosomes, except
557 on the microchromosome. We found the highest number of *CDSTR198* euchromatic
558 signals concentrated in chromosomes 2 and 5 (Figure 3a). Similar results were also
559 obtained by an overall analysis of 37 *CDSTR198* euchromatic arrays present in the *D.*
560 *buzzatti* assembled genome (Table S6). Interestingly, this analysis showed an equal
561 number of euchromatic arrays present on chromosomes 2 and 3 (11 arrays each),
562 followed by chromosomes 4 and 5 (six arrays each). The fewer euchromatic arrays
563 found in the *D. buzzatii* genome may result from the computational challenge of
564 repetitive element assembly (Treangen and Salzberg 2012), reinforcing the need of
565 hybridization experiments of satDNA families spread throughout euchromatin. In line
566 with this, it is relevant to suggest that some *CDSTR198* arrays identified by FISH may
567 be absent on assembled genomes. FISH on polytene chromosomes of *D. seriema*
568 showed *CDSTR198* located only in a few euchomatic sites (Figure 3b).

569 In contrast to transposable elements, satDNAs do not have the ability to
570 transpose by themselves. However, there are some reported examples showing that TEs
571 may act as a substrate for satDNA emergence and mobility (Dias et al. 2015; Mestrovic
572 et al. 2015; Satovic et al. 2016). We created a database containing the 500bp sequences
573 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

649 microchromosomes. According to Kuhn et al (2009), interspersion of repeats from non-
650 homologous satellites in the microchromosomes could be related to the peculiar
651 characteristics of these chromosomes, such as highly heterochromatic nature and low
652 content of genes, which could allow a more flexible interplay between repetitive
653 elements without deleterious effects.
654
655 **Differential transcription of cactophilic *Drosophila* satDNAs**
656 SatDNAs do not code for proteins and have been traditionally viewed as "junk
657 DNAs". However, there is a growing number of studies showing satDNA transcription
658 activity from yeast to mammals and the biological function of these transcripts has now
659 started to be appreciated. For example, satDNA transcripts were shown to be involved
660 in heterochromatin assembly, kinetochore formation and gene regulation (reviewed by
661 Biscotii et al. 2015; Ferreira et al. 2015). Moreover, transcription of satDNAs is usually
662 gender or stage specific and is often associated with differentiation and development
663 (Usakin et al. 2007; Pecinka et al. 2010).
664 Herein, we investigated whether the satDNAs that we analyzed are transcribed
665 by mapping the satDNA consensus sequences on the available RNA-seq data from *D.*
666 *buzzatii* and *D. mojavensis* (Guillen et al. 2015; Rius et al. 2016). Read counts were
667 calculated for embryos, third-staged larvae, pupae and for male and female adult
668 carcasses (Figure 8) (See methods).
669 Our analysis did not identify transcripts from the most abundant satDNAs in the
670 genome of *D. buzzatii* and *D. mojavensis*, *pBuM* and *CDSTR*130, respectively. As
671 discussed previously, both are the main candidates for centromeric function in these
672 species. This result was unexpected because previous studies in *Drosophila*
673 *melanogaster* showed that centromeric satellite RNAs in the form of long

674 polyadenylated products play an important role in the formation of the kinetochore
675 (Topp et al 2004; Chan et al. 2012; Rosic et al. 2014). However, our results do not
676 exclude the possibility that *pBuM* and *CDSTR*130 are transcribed. In this case, the
677 absence of satDNA transcripts may be related to the methodology used for RNA
678 extraction that preferentially captures poly(A) sequences. For example, satDNA
679 transcripts of *D. melanogaster* involve ncRNAs that do not have poly(A) tails (Usakin
680 et al. 2007).

681 Conversely, in all five analyzed tissues we detected transcripts derived from the
682 *CDSTR198* satDNA of *D. buzzatii* and from the *pBuM* satDNA of *D. mojavensis*. In
683 both cases, the transcripts were particularly abundant in tissues from pupae and males.
684 Interestingly, these two satDNAs are located in different genomic environments: while
685 *CDSTR198* arrays are located at several euchromatic loci (including some close to
686 genes; Table S7) in several *D. buzzatii* chromosomes, pBuM is exclusively located in
687 the heterochromatic microchromosome of *D. mojavensis*. Future studies will be needed
688 to address whether these transcripts participate in chromatin modulation and/or if they
689 affect the transcription of neighboring genes, as observed for satDNA transcripts of
690 *Drosophila* and other organisms (Menon et al. 2014; Fellicielo et al. 2015).

691

692

693

694

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

740 (identified by the yellow traced line). Each bar in the graphic represents a cluster of
741 similar reads. The *CDSTR130* and pBuM-1 satellite DNAs are indicated.

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

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

746 **Supplementary Tables Legends:**

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

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

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

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

754 **Table S5.** Description of the ten most abundant clusters of the *D. melanogaster* genome
755 identified by RepeatExplorer. The satDNA families with monomer lengths smaller than
756 50 bp are highlighted in bold.

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

759 **Table S7**. List of genes associated with *CDSTR198* arrays and their relative positions in
760 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*. We thank Prof. A. Bernardo Carvalho (Universidade Federal do Rio de Janeiro) for kindly sharing the *D. mojavensis* sequencing data with us. 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.

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

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

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

812 Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. 1999 Generation of a
813 widespread Drosophila inversion by a transposable element. *Science*, *285*(5426), 415-
814 418.

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

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

820 Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, et al.
821 2007 Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167),
822 203-218.

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

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

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

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

835 Dover, G. 1982 Molecular drive: a cohesive mode of species evolution. *Nature* 229
836 (5879): 111-117.

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

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

842 Ferreira, D., Meles, S., Escudeiro, A., Mendes-da-Silva, A., Adega, F., and Chaves, R.
843 2015. Satellite non-coding RNAs: the emerging players in cells, cellular pathways and
844 cancer. *Chromosome Research*, *23*(3), 479-493.

845 Fitzgerald, D. J., Dryden, G. L., Bronson, E. C., Williams, J. S., and Anderson, J. N.
846 1994 Conserved patterns of bending in satellite and nucleosome positioning
847 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.

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

893 Kuhn, G. C.S., Franco, F. F., Manfrin, M. H., Moreira-Filho, O., and Sene, F. M. 2007
894 Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single
895 pair of microchromosomes in species from the *Drosophila buzzatii*
896 cluster. *Chromosome research*, *15*(4), 457-470.

897 Kuhn, G. C.S., Sene, F. M., Moreira-Filho, O., Schwarzacher, T., and Heslop-Harrison,
898 J. S. 2008 Sequence analysis, chromosomal distribution and long-range organization
899 show that rapid turnover of new and old pBuM satellite DNA repeats leads to different
900 patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome*
901 *Research*, *16*(2), 307-324.

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

906 Kuhn, G.C,S. and Heslop-Harrison J.S. 2011. Characterization and genomic
907 organization of PERI, a repetitive DNA in the *Drosophila buzzatii* cluster related to
908 DINE-1 transposable elements and highly abundant in the sex chromosomes.
909 Cytogenetic and Genome Research 132:79–88

910 Kuhn, G. C.S., Küttler, H., Moreira-Filho, O., and Heslop-Harrison, J. S. 2012 The
911 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales
912 and association with genes. *Molecular biology and evolution*, 29:7-11.

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

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

918 Leung, W., Shaffer, C. D., Reed, L. K., Smith, S. T., Barshop, W., Dirkes, W., ... and
919 Yuan, H. 2015 Drosophila Muller F elements maintain a distinct set of genomic
920 properties over 40 million years of evolution. *G3: Genes| Genomes| Genetics*, *5*(5),
921 719-740.

922 Lohe, A. R., Hilliker, A. J., and Roberts, P. A. 1993 Mapping simple repeated DNA
923 sequences in heterochromatin of Drosophila melanogaster. *Genetics*, *134*(4), 1149-
924 1174.

925 Löpez, C. C., Nielsen, L., and Edström, J. E. 1996 Terminal long tandem repeats in
926 chromosomes form Chironomus pallidivittatus. *Molecular and cellular biology*, *16*(7),
927 3285-3290.

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

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

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

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

939 Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., et al.
940 2013 Comparative analysis of tandem repeats from hundreds of species reveals unique
941 insights into centromere evolution. *Genome biology*, *14*(1), R10.

942 Menon, D. U., Coarfa, C., Xiao, W., Gunaratne, P. H., and Meller, V. H. 2014 siRNAs
943 from an X-linked satellite repeat promote X-chromosome recognition in Drosophila
944 melanogaster. *Proceedings of the National Academy of Sciences*, *111*(46), 16460-
945 16465.

946 Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., and Plohl, M.
947 2015 Structural and functional liaisons between transposable elements and satellite
948 DNAs. *Chromosome research*, *23*(3), 583-596.

949 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008 Mapping
950 and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-
951 628.

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

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

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

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

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

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

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

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

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

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

981

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

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

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

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

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

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

997 Ruiz, A., Heed, W. B., and Wasserman, M. 1990. Evolution of the mojavensis cluster of
998 cactophilic Drosophila with descriptions of two new species. *Journal of Heredity*, *81*(1),
999 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).*

1022 Stephan, W., and Cho, S. 1994 Possible role of natural selection in the formation of
1023 tandem-repetitive noncoding DNA. *Genetics*, *136*(1), 333-341..

1024 Strachan, T., Webb, D., and Dover, G. A. 1985 Transition stages of molecular drive in
1025 multiple-copy DNA families in Drosophila. *The EMBO journal*, *4*(7), 1701.

1026 Sun, X., Wahlstrom, J., and Karpen, G. 1997 Molecular structure of a functional
1027 Drosophila centromere. *Cell*, *91*(7), 1007-1019.

1028 Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. 2003 Sequence analysis of a
1029 functional Drosophila centromere. *Genome research*, *13*(2), 182-194.

1030 Tamura, K., and Nei, M. 1993 Estimation of the number of nucleotide substitutions in
1031 the control region of mitochondrial DNA in humans and chimpanzees. *Molecular*
1032 *biology and evolution*, *10*(3), 512-526.

1033 Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011
1034 MEGA5: molecular evolutionary genetics analysis using maximum likelihood,
1035 evolutionary distance, and maximum parsimony methods. *Molecular biology and*
1036 *evolution*, *28*(10), 2731-2739.

1037 Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA
1038 sequences. In *DNA fingerprinting: State of the science* (pp. 21-28). Birkhäuser Basel.

1039 Treangen, T. J., and Salzberg, S. L. 2012 Repetitive DNA and next-generation
1040 sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1),
1041 36-46.

1042 Topp, C. N., Zhong, C. X., and Dawe, R. K. 2004 Centromere-encoded RNAs are
1043 integral components of the maize kinetochore. *Proceedings of the National Academy of*
1044 *Sciences of the United States of America*, *101*(45), 15986-15991.

1045 Urrego, R., Bernal-Ulloa, S. M., Chavarría, N. A., Herrera-Puerta, E., Lucas-Hahn, A.,
1046 et al. 2017 Satellite DNA methylation status and expression of selected genes in Bos
1047 indicus blastocysts produced in vivo and in vitro. *Zygote*, 1-10.

1048 Villasante, A., Abad, J. P., Planelló, R., Méndez-Lago, M., Celniker, S. E., and de
1049 Pablos, B. 2007 Drosophila telomeric retrotransposons derived from an ancestral
1050 element that was recruited to replace telomerase. *Genome research*, *17*(12), 1909-1918.

1051 Wei, K. H. C., Grenier, J. K., Barbash, D. A., and Clark, A. G. 2014 Correlated
1052 variation and population differentiation in satellite DNA abundance among lines of
1053 Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *111*(52),
1054 18793-18798.

1055 Yang, H. P., and Barbash, D. A. 2008 Abundant and species-specific DINE-1
1056 transposable elements in 12 Drosophila genomes. *Genome biology*, *9*(2), R39.

1057 Žinić, S. D., Ugarković, D., Cornudella, L., and Plohl, M. 2000 A novel interspersed
1058 type of organization of satellite DNAs in Tribolium madens
1059 heterochromatin. *Chromosome Research*, *8*(3), 201-212.

1060

1061

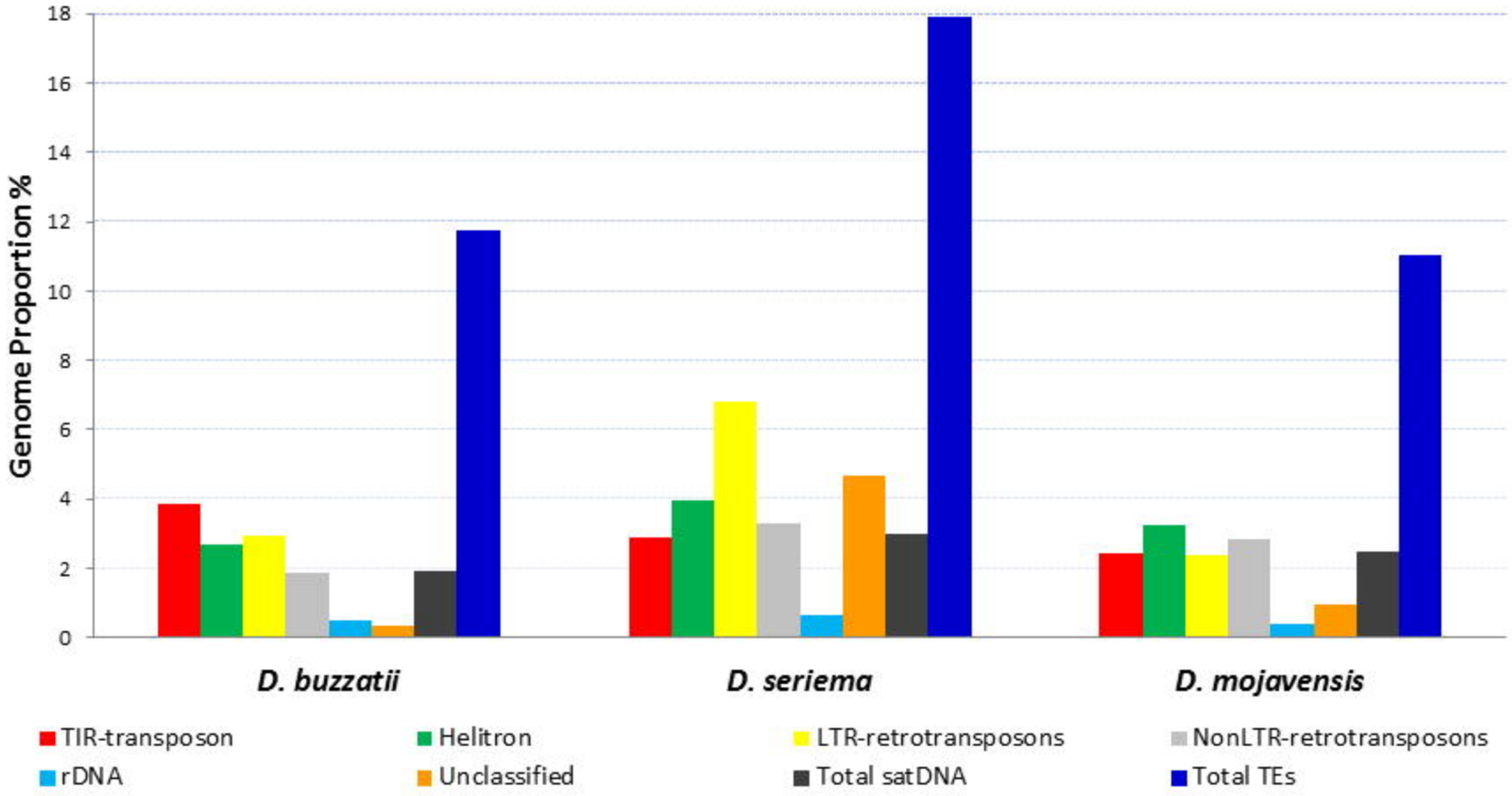
Genome Proportion %
18
16
14
12
10
8
6
4
2
0
D. buzzatii
D. seriema
D. mojavensis
TIR-transposon
Helitron
LTR-retrotransposons
NonLTR-retrotransposons
rDNA
Unclassified
Total satDNA
Total TEs

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:

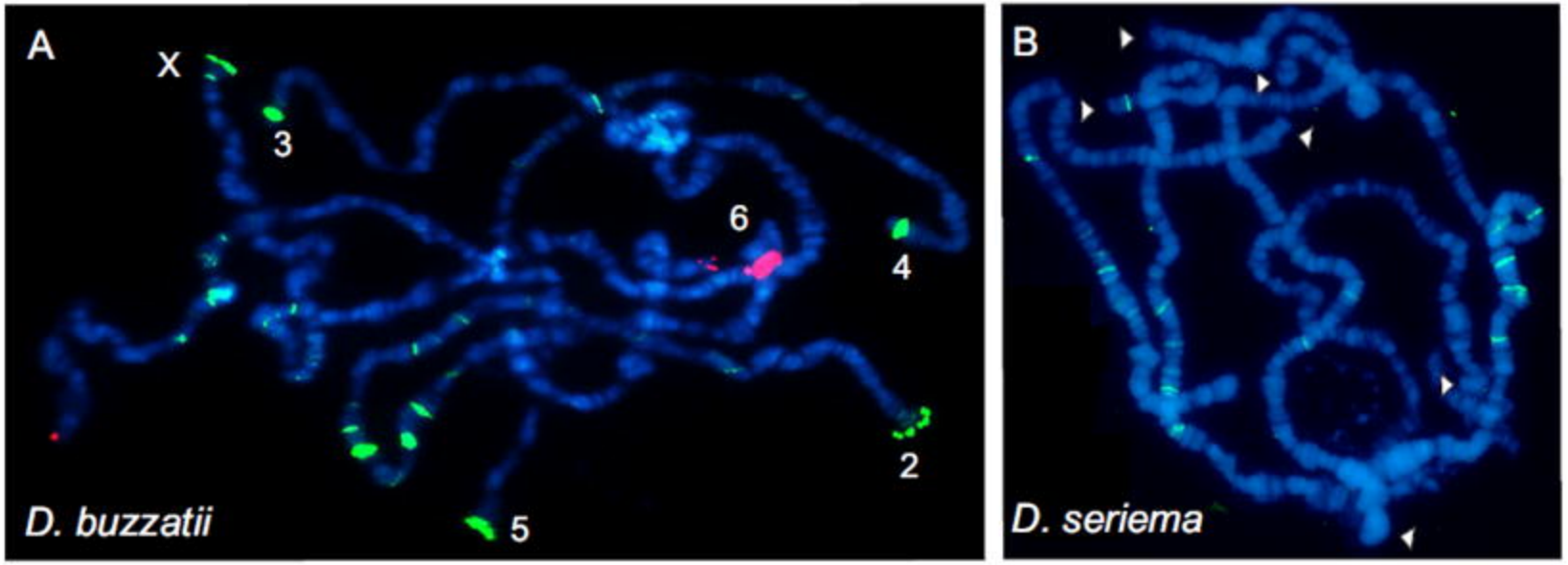

Fig. 3: FISH on polytene chromosomes of *D. buzzatii* (A) and (B) *D. seriema* using satDNA probes for *pBuM* (red) and *CDSTR198* (green) (Arrowheads indicate telomeric regions).

Figure 4.

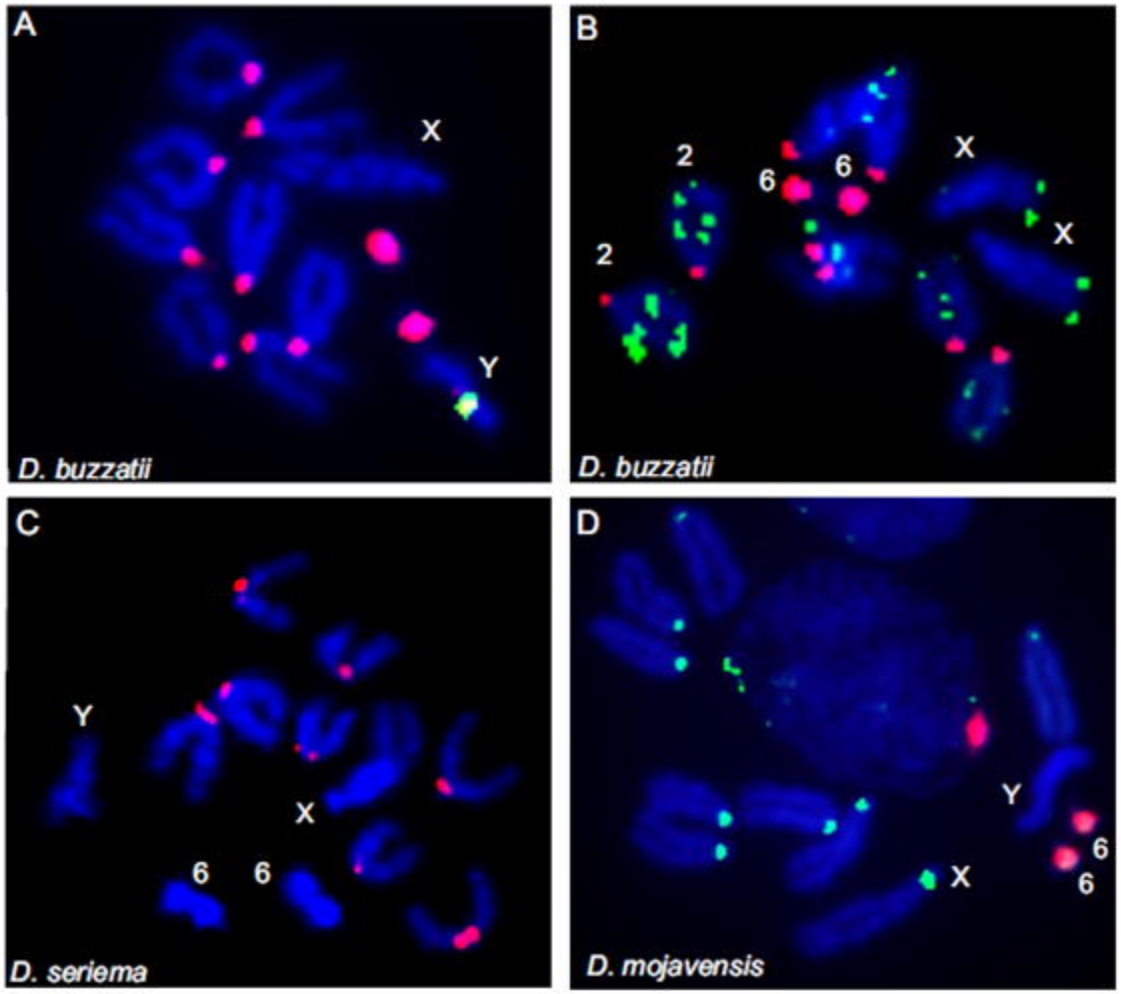

Fig. 4: FISH on mitotic chromosomes using satellite DNA probes. (A) *pBuM-1a* (red) and pBuM-1b (green) satDNA probes on *D. buzzatii;* **B.** *pBuM-1a* (red) and *CDSTR198* (green) probes on *D. buzzatii;* **C.** *CDSTR138* (red) on *D. seriema* **(D)** *CDSTR130* (green) and *pBuM* (red) probes on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

Figure 5

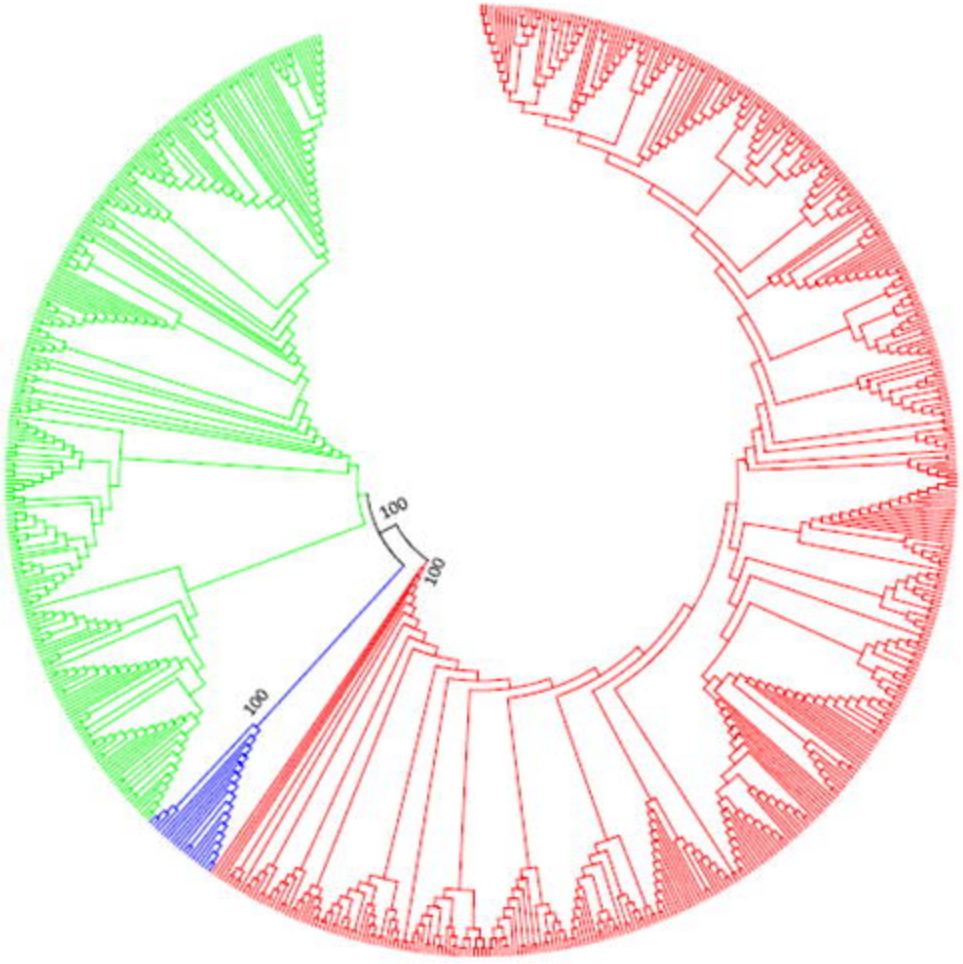

Fig. 5: NJ tree containing a sample of *pBuM* repeats extracted from the sequenced genomes of *Drosophila buzzatii* (green), *D. seriema* (blue) and *D. mojavensis* (red). The tree was estimated using the T93 substitution model with 1,000 bootstrap replicas.

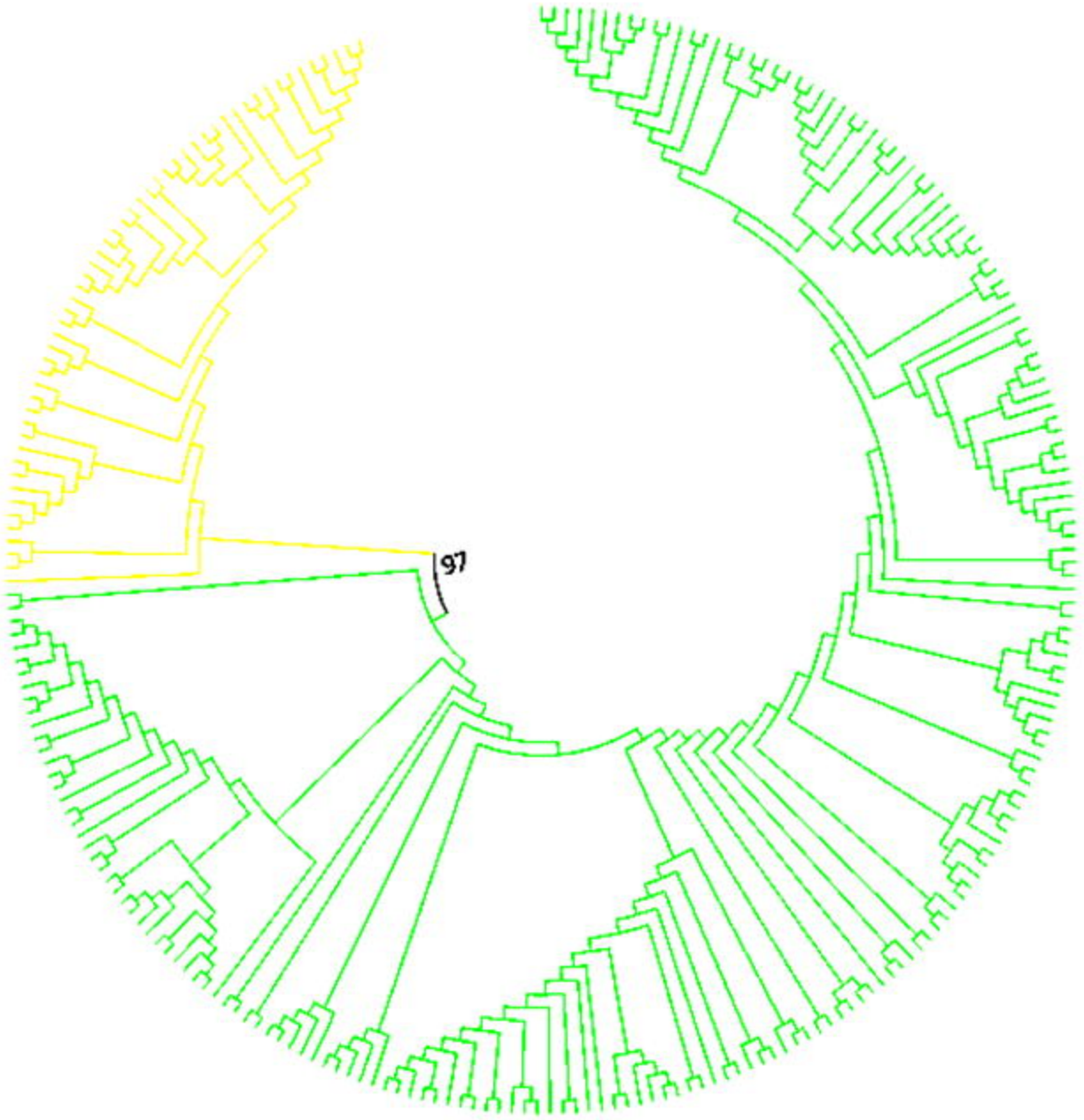
97

Figure 7.

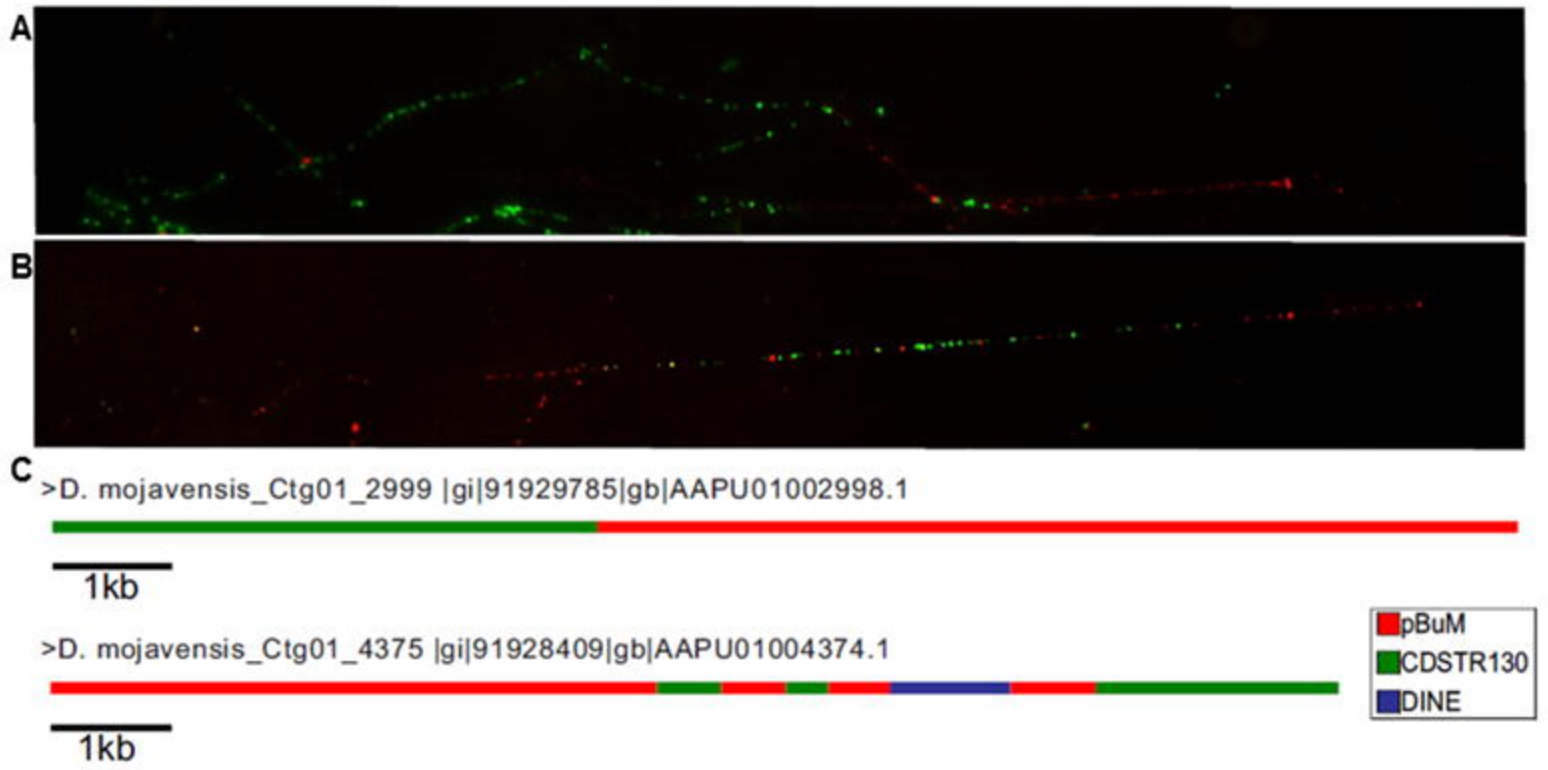

Fig. 7: A-B FISH with *CDSTR*130 (green) and pBuM (red) probes onto extended DNA fibers of *D. mojavensis*. (C) Schematic representation of *CDSTR130* and *pBuM* organization found on *contigs Ctg01_2999(*AAPU01002998.1) and *Ctg01_4375* (AAPU01004374.1) retrieved from the *D. mojavensis* assembled genome.

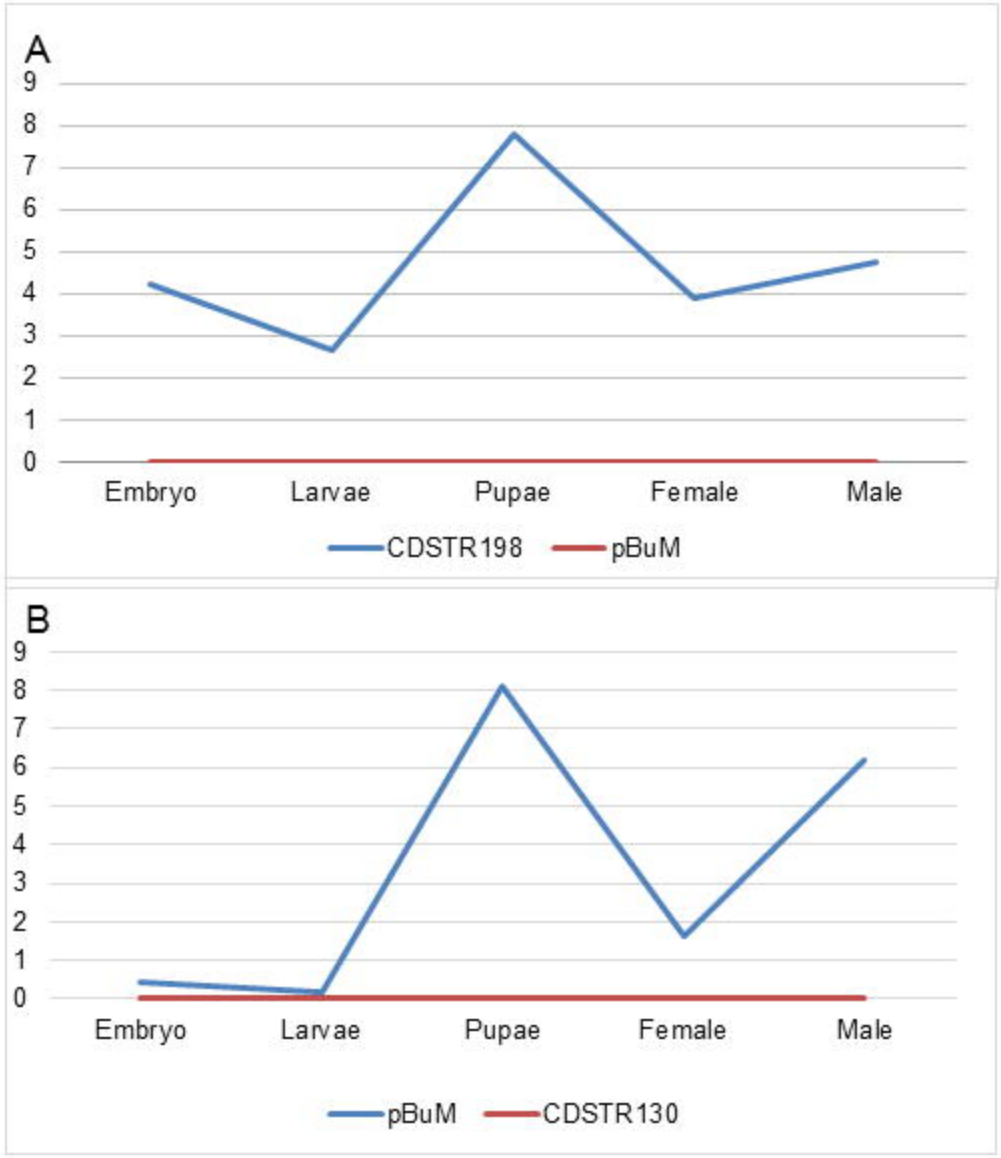
A
9
8
7
6
5
4
3
2
1
0
Embryo
Larvae
Pupae
Female
Male
CDSTR198
pBuM
B
9
8
7
6
5
4
3
2
1
0
Embryo
Larvae
Pupae
Female
Male
pBuM
CDSTR130

Figure 9:

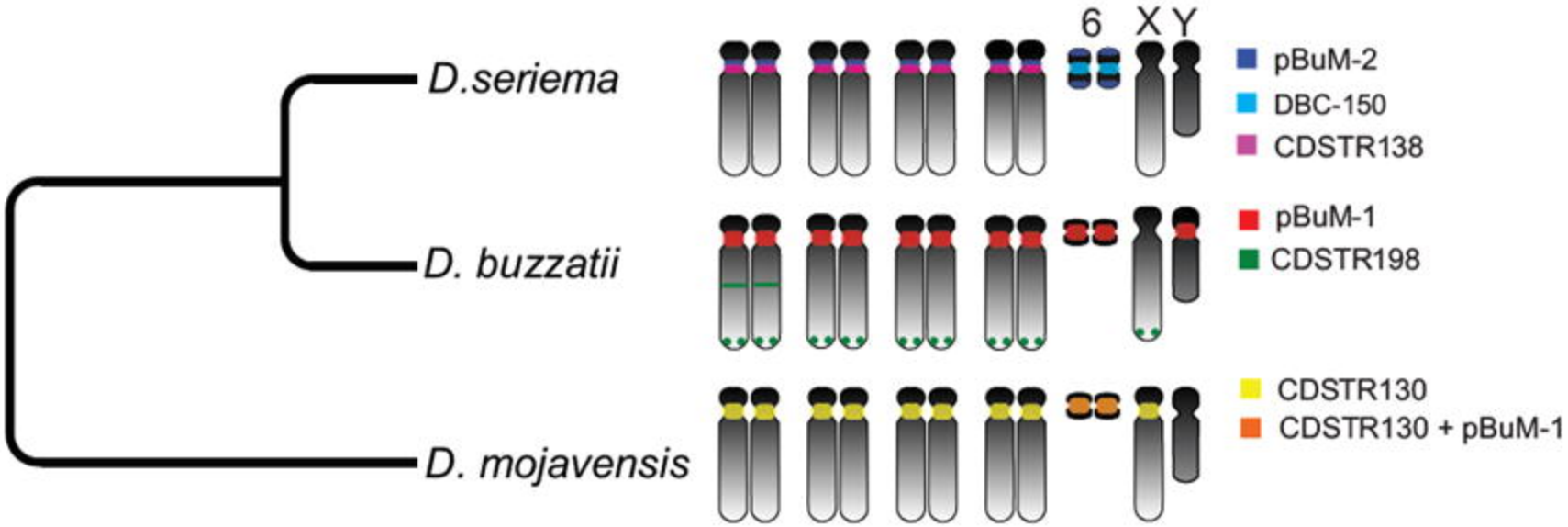

Fig. 9: Representative ideogram showing the chromosomal localization of all satDNAs identified in *D. buzzatii*, *D. seriema* and *D. mojavensis*.

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).