In-Depth Satellitome Analyses of 37 *Drosophila* Species Illuminate Repetitive DNA Evolution in the *Drosophila* Genus

Leonardo G. de Lima [1,*] and Francisco J. Ruiz-Ruano [2,3]

[1] Stowers Institute for Medical Research, 1000 E 50th Street, Kansas City, MO 64110, USA

[2] Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3] School of Biological Sciences, Norwich Research Park University of East Anglia, Norwich, UK

*Corresponding author: E-mail: lgomesdelima@stowers.org.

Accepted: 27 April 2022

Abstract

Satellite DNAs (SatDNA) are ubiquitously present in eukaryotic genomes and have been recently associated with several biological roles. Understanding the evolution and significance of SatDNA requires an extensive comparison across multiple phylogenetic depths. We combined the *RepeatExplorer* pipeline and cytogenetic approaches to conduct a comprehensive identification and analysis of the satellitome in 37 species from the genus *Drosophila*. We identified 188 SatDNA-like families, 112 of them being characterized for the first time. Repeat analysis within a phylogenetic framework has revealed the deeply divergent nature of SatDNA sequences in the *Drosophila* genus. The SatDNA content varied from 0.54% of the *D. arizonae* genome to 38.8% of the *D. albomicans* genome, with the SatDNA content often following a phylogenetic signal. Monomer size and guanine–cytosine-content also showed extreme variation ranging 2–570 bp and 9.1–71.4%, respectively. SatDNA families are shared among closely related species, consistent with the SatDNA library hypothesis. However, we uncovered the emergence of species-specific SatDNA families through amplification of unique or low abundant sequences in a lineage. Finally, we found that genome sizes of the *Sophophora* subgenus are positively correlated with transposable element content, whereas genome size in the *Drosophila* subgenus is positively correlated with SatDNA. This finding indicates genome size could be driven by different categories of repetitive elements in each subgenus. Altogether, we conducted the most comprehensive satellitome analysis in *Drosophila* from a phylogenetic perspective and generated the largest catalog of SatDNA sequences to date, enabling future discoveries in SatDNA evolution and *Drosophila* genome architecture.

Key words: satellite DNA, *Drosophila*, repetitive DNA, genome size evolution, *RepeatExplorer*.

Significance

Satellite DNAs are large arrays of tandemly repeated sequences that represent a large portion of eukaryotic genomes and are associated with a variety of mechanisms that shape genome architecture and evolution. Here, we used a broad phylogenetic survey approach associating next-generation sequencing data and cytogenomics to generate the largest satellite DNAs (SatDNA) sequences analysis in *Drosophila* to date. We used the *RepeatExplorer* pipeline to generate a catalog of 188 SatDNA sequences in 37 species of *Drosophila* with 112 of them being newly characterized, resulting in the most comprehensive satellitome analysis in *Drosophila* species so far. Our findings indicate that *Drosophila* subgenus genomes are strongly shaped by SatDNA expansion/contractions, whereas *Sophophora* species seem to hold a stronger correlation with transposable elements (TE) content variation, suggesting that each subgenus genome size evolution can be differently influenced by the modulation of SatDNA and TE contents.

Introduction

Virtually all eukaryotic genomes studied to date contain tandem arrays in which the basic units or monomers are repeated in a head-to-tail fashion known as satellite DNAs (SatDNA) (Charlesworth et al. 1994). SatDNA can comprise up to 50% of eukaryotic genomes and are usually found in long tandemly repeated arrays that can span up to megabases (reviewed by Schmidt and Heslop-Harrison 1998; Plohl et al. 2012). SatDNA abundance can differ dramatically among species, appearing to evolve by array expansion and shrinkage of related repeat variants (Nijman and Lenstra 2001; Slamovits et al. 2001). In *Drosophila* species, SatDNA can account for more than 30% of the genome, and amplification/contraction events of distinct SatDNA families have been identified as a crucial factor in shaping the architecture and size of the *Drosophila* genome (Bosco et al. 2007). Moreover, species evolution may be associated with SatDNA given that rapid changes in copy number can trigger rapid genome changes (Dover and Tautz 1986; Gregory and Johnston 2008; Ferree and Barbash 2009; Brand and Levine 2021).

SatDNA are often associated with heterochromatin and the low gene content in these regions led to the misconception that SatDNA have no essential function, being labeled as "junk" DNA (Ohno 1972). However, it is now clear that SatDNA sequences can regulate cellular processes such as kinetochore assembly, X chromosome recognition, and meiotic chromosome segregation (Dernburg et al. 1996; Kuhn 2015; Shatskikh et al. 2020). For example, recent publications have demonstrated that the *1.688* SatDNA family in *D. melanogaster* is crucial for centromeric function, chromosome missegregation in hybrids, dosage compensation, and heterochromatin formation (Cattani and Presgraves 2012; Ferree and Prasad 2012; Menon et al. 2014; Rošić et al. 2014). These data have strengthened the argument that SatDNA can play a significant role in genomic organization and species adaptation.

Despite their importance for genome organization, function, and evolution, SatDNA have been historically neglected in the *Drosophila* genomes and eukaryotes in general. SatDNA have been mainly utilized as taxonomic markers based on the number and morphology of signal-bearing chromosomes and on the localization of signals in different species (Picariello et al. 2002; Bueno et al. 2021; Cabral-de-Mello et al. 2021), but this approach usually requires a priori knowledge of its sequences. In *Drosophila,* SatDNA investigations began later than in other animals such as frogs and mice, starting with the use of CsCl density gradients to characterize the most abundant sequences in *D. virilis*, *D. melanogaster*, and *D. hydei* genomes (Laird and McCarthy 1968; Gall et al. 1971; Gall and Atherton 1974; Barnes et al. 1978; Renkawitz 1979). During the last decades, SatDNA have been studied from a small sample of cloned repeats obtained by sequence biased methodological approaches (usually 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; Kuhn et al. 2008).

The development of next-generation sequencing (NGS) technologies more than a decade ago has generated a vast body of sequencing data and new tools for genome assembly. However, due to their repetitive natures, SatDNA sequences create ambiguities when aligning and assembling NGS data (Treangen and Salzberg 2012). As a result, recent genome reports lack a curated characterization of SatDNA sequences as they rely on the repetitive sequences deposited in the nucleotide databases. Indeed, despite the large repertoire of sequenced genomes available for *Drosophila* species, only 59 SatDNA families from *Drosophila* are deposited in the Genbank and Repbase databases (last accessed July 2021; supplementary material, Supplementary Material online). The recent availability of the cluster-based de novo identification pipeline *RepeatExplorer* (Novák et al. 2013) allowed us to comprehensively identify an extensive collection of repetitive sequences from a given genome using large NGS datasets, fostering the characterization of new SatDNA (de Lima et al. 2017; Silva et al. 2019). Despite these advances, recent studies have focused on SatDNA dynamics in restricted phylogenetic groups such as the *D. melanogaster* subgroup or *D. virilis* subgroup (Dias et al. 2014; Larracuente 2014; Khost et al. 2017; Silva et al. 2019; de Lima et al. 2020; Sproul et al. 2020). An expanded phylogenetic analysis beyond these two subgroups would benefit our understanding of genome evolution. Thus, a representative catalog of the SatDNA families and a comprehensive approach are important steps toward understanding their evolutionary pathways and how they are linked to genome evolution.

We performed here a high-throughput analysis of the satellitome from 37 *Drosophila* species and their correlation with genome size. We confirmed the presence of the previously described SatDNA and identified 112 new SatDNA sequences present in *Drosophila* genomes. Moreover, we compared the maintenance of SatDNA families throughout *Drosophila* phylogeny and provide evidence for species-specific birth and amplification of new SatDNA families in the context of genome size. Altogether, our work is the most comprehensive comparison of SatDNA in *Drosophila* species to date that is focused on a phylogenetic perspective to generate a large database that will improve *Drosophila* genome annotation.

Results

A De Novo and In Silico Identification of SatDNA-like Families Reveals Variation in the *Drosophila* Satellitome

We performed a de novo identification of repeats in 37 *Drosophila* species with the *RepeatExplorer* pipeline based on raw Illumina reads. We chose this approach to ensure the comprehensive characterization of SatDNA and other repetitive families, because methods based on assembled genomes hold potential biases due to the difficulty of assembling highly repetitive regions (*Drosophila* 12 Genomes Consortium 2007; Guillén et al. 2015; Kim et al. 2021). The overall de novo clustering of reads used at least 0.35-fold of the genome coverage (supplementary table S1, Supplementary Material online). This genome coverage is sufficient to identify the most repetitive elements in eukaryotic genomes (Novák et al. 2017; Fu et al. 2019; Silva et al. 2019) (see Materials and Methods). The present study expands more than 3-fold the available SatDNA dataset from 59 to 188 SatDNA-like sequences of *Drosophila*, including new SatDNA in several species and previously described SatDNA in additional species (fig. 1). The overall characterization of all SatDNA sequences and the overall repetitive DNA content for 37 species are given in figure 2 and table 1; supplementary table S2, Supplementary Material online. All consensus sequences for the SatDNA families obtained in this study are present in supplementary material S1, Supplementary Material online. We considered only clusters with genomic proportion equal to or higher than 0.01%. Furthermore, all SatDNA-like monomers were independently identified by the *RepeatProfiler* pipeline (Negm et al. 2020) to confirm the expected coverage depth profiles for each SatDNA-like family (supplementary material S2, Supplementary Material online). The high number of SatDNA families found in the genomes of the *Drosophila* species corroborates the assumption that eukaryote genomes usually contain a high diversity of SatDNA families (Melters et al. 2013). The overall repetitive DNA (i.e., all repetitive sequences above the 0.01% cut-off) comprised anywhere from 8.5% of a genome (*D. erecta*) to 48.1% of a genome (*D. albomicans* males) (fig. 2 and table 1). We assigned clusters from all species to specific repeat types and families (supplementary material S4, Supplementary Material online), except for the *D. seriema* genome which showed ~4% bacterial contamination and was removed by manual curation.

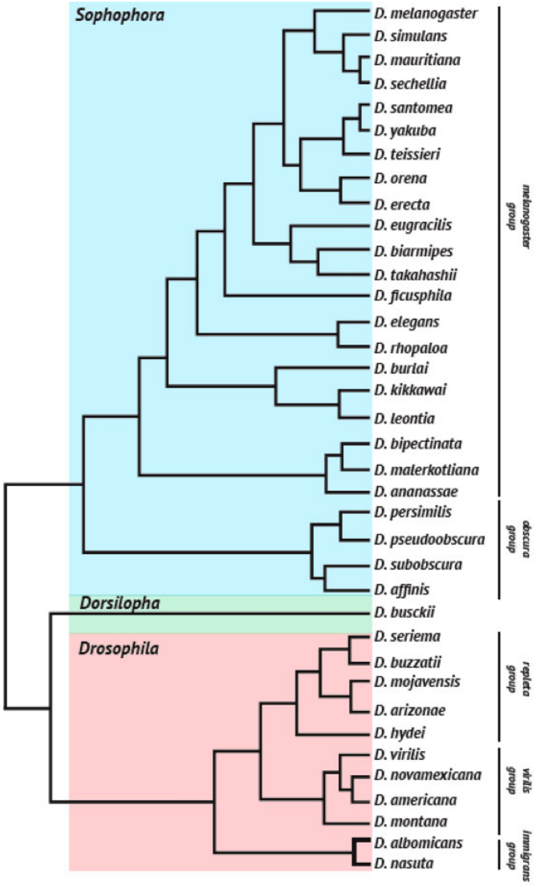


Fig. 1.—Phylogenetic relationship among 37 *Drosophila* species analyzed in the present study. The species are presented according to the phylogenetic tree proposed by Russo et al. (2013) and are differentially colored by respective subgenus.

In general, *Drosophila* SatDNA differ in nucleotide sequence, complexity, motif length, abundance, and chromosome localization. SatDNA are roughly classified by motif length as simple or complex, with simple SatDNA having a motif length of 5–12 bp, whereas complex repeats are ~150–360 bp (reviewed by Plohl et al. 2012). The 188 SatDNA-like families identified in the present work showed high variation for monomer length (2–570 bp) (fig. 3*A*). We describe a trimodal distribution of motif length with a higher prevalence of SatDNA monomers <50 bp (54), followed by 180–200 bp monomers (33). Although many different simple SatDNA families were identified, a major limitation of the *RepeatExplorer* pipeline is that high similarity among simple SatDNA might mistakenly allow clustering of unique simple SatDNA families. Simple SatDNA are known to comprise the heterochromatic regions of *D. melanogaster* and *D. virilis* subgroups species. Each of these simple SatDNA share 80–90% sequence similarity inside their respective subgroup (Jagannathan et al. 2017; Flynn et al. 2020). The *D. melanogaster* genome is rich in SatDNA families, most of them being (12) 7 bp or less (Lohe et al. 1993, reviewed in Lauria-Sneideman and Meller 2021). Seven out of 12 families share a similar nucleotide composition (AANAB) and are known to form long interspersed arrays (Chang and Larracuente 2017; Chang et al. 2019). Due to the clustering cut-off parameters available in *RepeatExplorer* (Novák et al. 2013, see Materials and Methods), simple SatDNA families in *D. melanogaster* are under-identified in a number of different families. However, a detailed examination of the assembled contigs generated by the *RepeatExplorer* pipeline shows that similar nucleotide

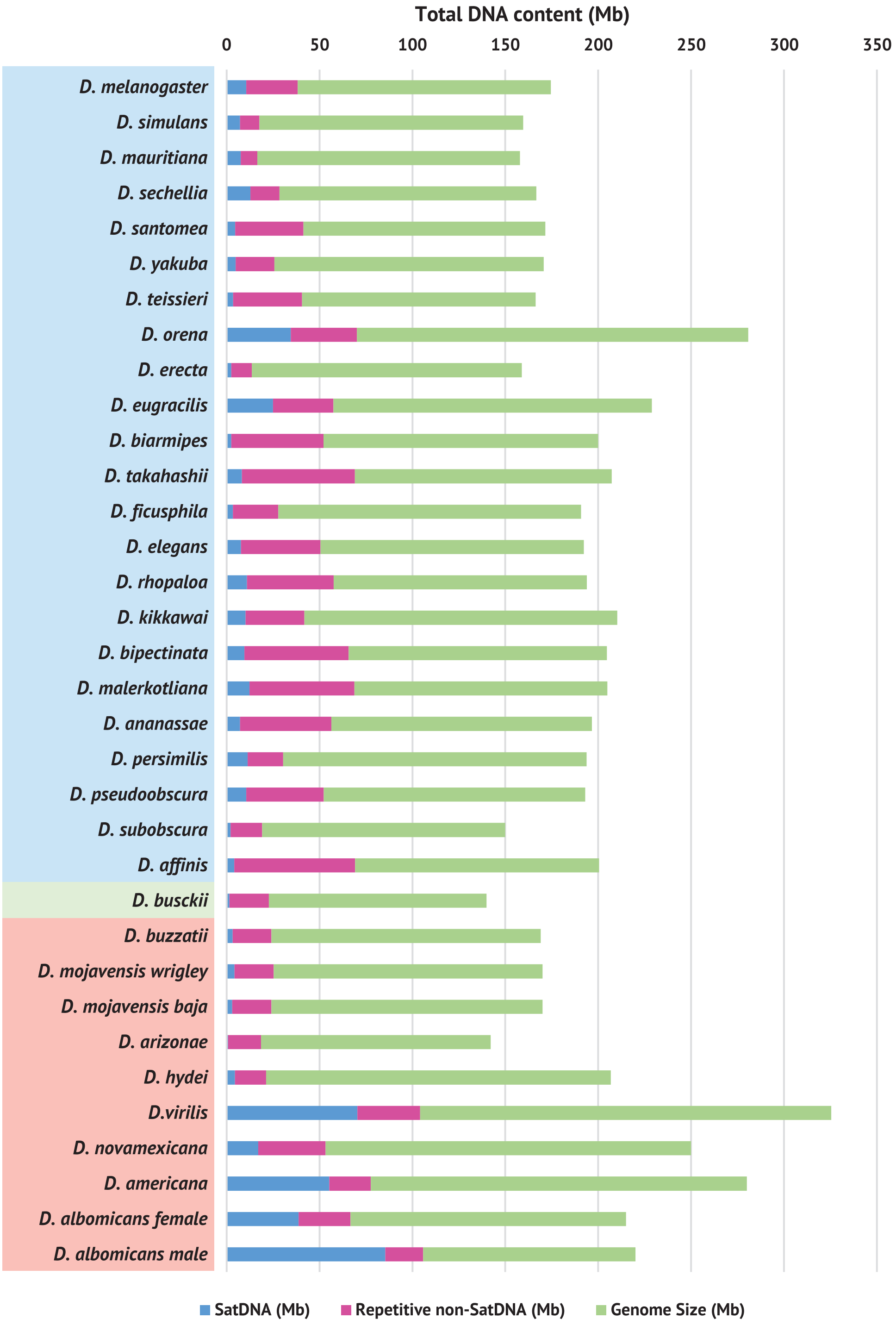


Fig. 2.—Genome size distribution and repeat composition of 32 *Drosophila* species. The species are plotted according to the phylogenetic tree topology proposed by Russo et al. (2013). The respective genome sizes (green), the non-SatDNA repetitive content (magenta), and SatDNA content (blue) are shown in the total amount of DNA comprised by each feature. *D. burlai*, *D. leontia*, *D. montana*, *D. nasuta*, and *D. seriema* data are not plotted due to the absence of genome size estimates for these species. One subspecies of *D. mojavensis* and both male and female data of *D. albomicans* are plotted. Genome size estimations are from the Animal Genome Size Database (www.genomesize.com), Bosco et al. (2007), Hjelmen and Johnston (2017), and Hjelmen et al. (2019).

composition SatDNA families collapsed in the same cluster (supplementary table S3, Supplementary Material online). If expanded to all datasets, it is expected that the overall amount of SatDNA observed might be similar but the total number of short SatDNA families is underrepresented, indicating that k-mer approaches might be necessary to

Table 1

Repetitive Content Estimation and Satellite DNA Contribution of the 37 *Drosophila* Species Included in This Study

Species	Subgenus	Repetitive Content%	SatDNA Content%	Genome Size (Mb)
D. melanogaster	*Sophophora*	21.9	6.6	174.5
D. simulans	*Sophophora*	11.1	4.53	159.6
D. mauritiana	*Sophophora*	10.5	4.86	157.9
D. sechellia	*Sophophora*	17.1	7.72	166.7
D. santomea	*Sophophora*	24.1	2.7	171.5
D. yakuba	*Sophophora*	15.1	2.83	170.7
D. teissieri	*Sophophora*	24.4	2.09	166.3
D. orena	*Sophophora*	25	12.31	280.7
D. erecta	*Sophophora*	8.5	1.62	158.9
D. eugracilis	*Sophophora*	25.1	10.89	228.9
D. biarmipes	*Sophophora*	26.1	1.27	200
D. takahashii	*Sophophora*	33.3	3.95	207.3
D. ficusphila	*Sophophora*	14.6	1.76	190.8
D. elegans	*Sophophora*	26.3	4.01	192.2
D. rhopaloa	*Sophophora*	29.7	4.67	193.9
D. burlai	*Sophophora*	24.8	3.12	N/A
D. kikkawai	*Sophophora*	19.9	4.85	210.2
D. leontia	*Sophophora*	15.2	1.81	N/A
D. bipectinata	*Sophophora*	32.1	4.72	204.6
D. malerkotliana	*Sophophora*	33.6	6.04	204.9
D. ananassae	*Sophophora*	28.7	3.68	196.6
D. persimilis	*Sophophora*	15.7	5.87	193.7
D. pseudoobscura	*Sophophora*	27.1	5.48	193
D. subobscura	*Sophophora*	12.8	1.4	150
D. affinis	*Sophophora*	34.5	2.07	200.5
D. busckii	*Dorsilopha*	16.3	1.1	139.9
D. seriema	*Drosophila*	26.2	2.9	N/A
D. buzzatii	*Drosophila*	14.3	1.9	169
D. mojavensis wrigley	*Drosophila*	14.9	2.49	170
D. mojavensis baja	*Drosophila*	14.2	1.76	170
D. arizonae	*Drosophila*	13	0.54	142.1
D. hydei	*Drosophila*	10.3	2.16	206.8
D.virilis	*Drosophila*	32	21.63	325.4
D. novamexicana	*Drosophila*	21.3	6.82	250
D. americana	*Drosophila*	27.7	19.75	280
D. montana	*Drosophila*	39	27.41	N/A
D. albomicans female	*Drosophila*	31	18.1	215
D. albomicans male	*Drosophila*	48.1	38.8	220
D.nasuta female	*Drosophila*	22	16.5	N/A
D.nasuta male	*Drosophila*	42	33.93	N/A

identify arrays of simple SatDNA shorter than 10 bp (Wei et al. 2018). Notably, four nonhomologous simple SatDNA families (with monomers from 2 to 14 bp) lacking sequence similarities were correctly characterized by our methodology in the *D. hydei* genome when compared with previous analyses (Burgtorf and Bünemann 1994). Further, we did not identify any new SatDNA family organized in multi-unit higher-order repeats, although we confirmed the previously described *pBuM-2* α/β alternating repeats of *D. seriema* (Kuhn et al. 2008; de Lima et al. 2017). However, future long-read sequencing approaches are necessary for a proper study of higher-order repeats (Kunyavskaya et al. 2021).

Our analysis showed that the guanine–cytosine (GC) content of SatDNA sequences in *Drosophila* varies widely from 9.1% to 71.8%. We observed that SatDNA sequences tend to be AT-rich, although some SatDNA sequences are very GC-rich (fig. 3*B*). AT-rich SatDNA are common in insect genomes and are suggested to contribute to the duplex curvature that enhances nucleosome stability (reviewed in Palomeque and Lorite 2008). Contrastingly, higher GC content sequences are suggested to elevate double-strand breaks, base substitutions, and a high rate of deletions by DNA polymerase slippage in eukaryotes (Kiktev et al. 2018). Short-read sequencing tends to be biased toward higher GC content sequences due to PCR-based

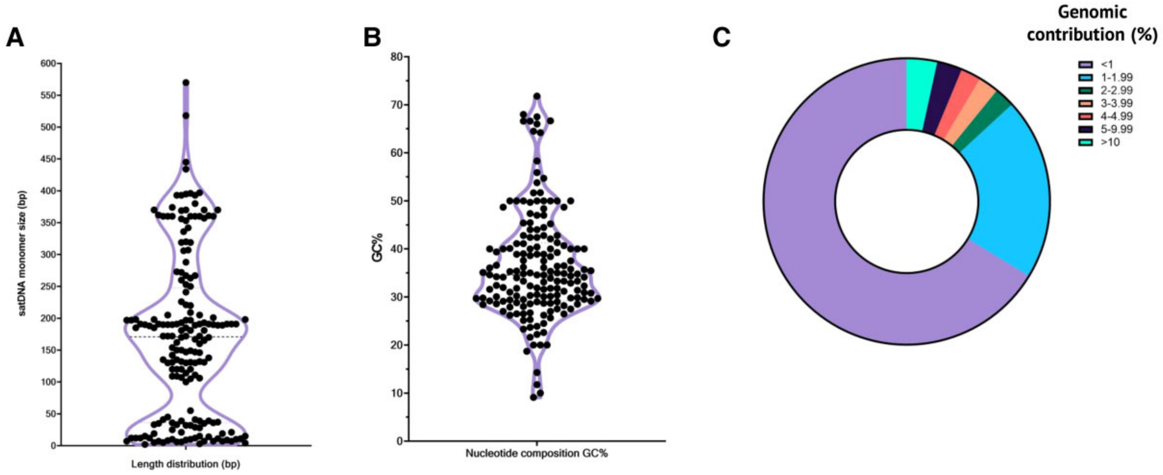


Fig. 3.—Main features of 188 SatDNA families characterized by size, GC content, and abundance of SatDNA in the *Drosophila* genus. (*A*) Monomer size plot for all 188 SatDNA indicates a trimodal distribution. (*B*) Distribution of the GC content for all 188 SatDNA identified indicates a significant differential pattern of sequence composition. (*C*) Genomic contribution for all 188 SatDNA identified in the *Drosophila* genus. We scored how many make up <1% of their resident species genome (<1%), how many make up between 1% and 2%, etc. Most SatDNA have low abundance (<1%) and comprise <1% of their resident genome (purple).

amplification biases in the libraries. However, the previously annotated SatDNA sequences had similar GC content compared with the same SatDNA sequence identified with *RepeatExplorer*. *Drosophila* genomes are, in general, AT-rich (*Drosophila* 12 Genomes Consortium 2007), suggesting that SatDNA sequences may share the overall base composition of their resident genomes.

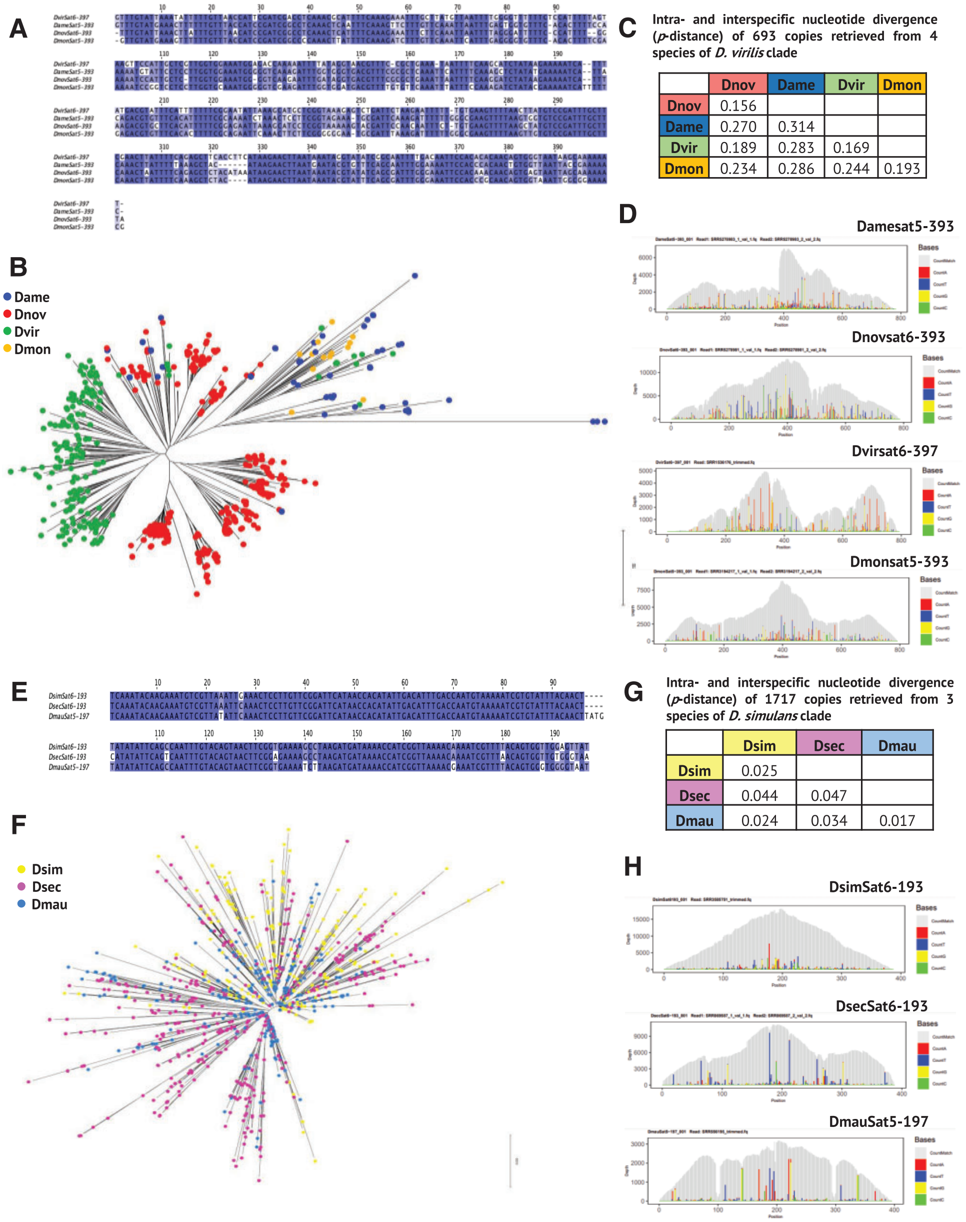
In Silico Characterization Reveals New SatDNA Families in *Drosophila*

We identified a total of 188 SatDNA-like sequences in 37 species of *Drosophila*, 112 of them corresponding to previously undescribed SatDNA-like families distinct from sequences present in GenBank, RepBase, and SatDNA literature records (supplementary material S1, Supplementary Material online). Importantly, we characterized new SatDNA-like sequences in species without previous SatDNA identification efforts (e.g., *D. burlai*, *D. leontia*, *D. malerkotliana*, and *D. nasuta*). Other species had prior limited analysis of SatDNA, such as Strachan et al. (1985) and de Lima et al. (2020) that focused only on the *1.688* SatDNA family, or Melters et al. (2013) that focused only on the most abundant repetitive DNA in each species. New SatDNA-like sequences were also discovered in species with previous SatDNA annotations. Notably, new SatDNA-like sequences were characterized even in well-characterized species, such as the *D. virilis subgroup and simulans* clade species. For instance, a new low abundant SatDNA-like family with 393–397 bp monomers is shared among *D. virilis, D. americana, D. montana, and D. novamexicana (DvirSat6-397, DameSat5-393, DmonSat5-393, and DnovSat6-393*, respectively) (fig. 4*A* and *B*). These SatDNA-like families can also be found flanked by previously described heterochromatic SatDNA DvirSat1-7 arrays in *D. virilis* (Gall et al. 1971; Silva et al. 2019), or other SatDNA families in *D. novamexicana* (DnovSat5-33; DnovSat4-190) suggesting it has heterochromatic localization in these species (e.g., VNHH02000100.1; VNHH02000181.1; BJEM01000096.1). These SatDNA-like families revealed an initial pattern of concerted evolution, especially between *D. virilis and D. novamexicana*, demonstrated by species-specific clusters along with lower intraspecific nucleotide diverge when compared with interspecific values (fig. 4).

In addition to identifying new SatDNA families, our approach confirmed a recently identified 193 bp SatDNA-like family restricted to the *simulans clade* species X chromosome (Chakraborty et al. 2021) (fig. 4*C* and *D*; supplementary table S1 and material S1, Supplementary Material online). The 193–7 bp SatDNA copies of the *simulans* clade do not indicate clear intraspecific phylogenetic clusters and show similar interspecific and intraspecific nucleotide diverges values (fig. 4). Our results imply that SatDNA sequences are still underrepresented in databases and genome assemblies. Future cytogenetics and long-read-based analyses are required to uncover the complete chromosomal and genomic organizations of these newly described SatDNA families.

Transposable Elements Structural Sequences Display a Minor Influence on SatDNA Origins in *Drosophila*

A clear relationship between TE and the origin, amplification, and homogenization of SatDNA has been debated in



Intra- and interspecific nucleotide divergence (*p*-distance) of 693 copies retrieved from 4 species of *D. virilis* clade

	Dnov	Dame	Dvir	Dmon
Dnov	0.156			
Dame	0.270	0.314		
Dvir	0.189	0.283	0.169	
Dmon	0.234	0.286	0.244	0.193

Intra- and interspecific nucleotide divergence (*p*-distance) of 1717 copies retrieved from 3 species of *D. simulans* clade

	Dsim	Dsec	Dmau
Dsim	0.025		
Dsec	0.044	0.047	
Dmau	0.024	0.034	0.017

Fig. 4.—Consensus sequence alignments, phylogenetic reconstruction trees, and nucleotide divergence of SatDNA families in *D. virilis* group and *simulans* clade. (*A* and *E*) Monomeric consensus sequence alignment of a new 393–7 and 193–7 bp SatDNA family shared among species from *D. virilis* group and *simulans* clade species, respectively. (*B*) Unrooted Neighbor-Joining phylogenetic tree using 635 full-length 393–7 bp SatDNA-like monomers derived from *D.virilis*, *D. americana*, *D. novamexicana*, and *D. montana* suggesting an initial level of species-specific homogenization. (*C*) Intraspecific and interspecific nucleotide divergence (*p*-distance) of 393–7 bp SatDNA in the *D. virilis* group support the hypothesis of initial stages of concerted evolution in *D. virilis* and *D. novamexicana* genomes. (*F*) Unrooted Neighbor-Joining phylogenetic tree using 1,717 full-length 193–7 bp SatDNA-like monomers derived from *D. simulans*, *D. sechellia*, and *D. mauritiana* species. (*G*) Intraspecific and interspecific nucleotide divergence (*p*-distance) of *simulans* clade 193–7 bp SatDNA sequences suggests a low level of species-specific homogenization. (*D* and *H*) Variant-enhanced repeat profiles from the putative new 393–7 and 193–7 bp SatDNA families in *D. virilis* group and *simulans* clade species, respectively.

several insect species, especially in *Drosophila* (Palomeque and Lorite 2008; McGurk and Barbash 2018). Herein we identified three previously described tandemly repeated families present in the *D. virilis* group that originated from TEs: *pvB370* SatDNA, *Tetris-220*, and *DINE-TR1 CTRs* (Heikkinen et al. 1995; Dias et al. 2014, 2015, respectively). We identified new SatDNA families that have been annotated as if they share sequence similarity to TEs, two in the *D. biarmipes* genome and one in the *D. pseudoobscura* group species (*D. pseudoobscura* and *D. persimilis*). The first one, DbiaSat2-263, was deposited in *RepBase* as *Copia-2_DTa-I_1p*, whereas DbiaSat5-215 was present as *Gypsy-11_DEu-I*. Despite the annotations, we were unable to find sequence similarity between *DbiaSat2-263* and *Dbia5-215* and the *Copia* and *Gypsy* TE families, respectively. We further identified that the *Gypsy18-I_Dpse* sequence contains a 228 bp LTR sequence, indicating an incorrect annotation of this specific TE sequence. The *Gypsy18-LTR_Dpse* sequence deposited in the *RepBase* database is a large array of a 21 bp SatDNA family shared among *D. pseudoobscura* and *D. persimilis.* Further, to confirm our characterization of DbiaSat2-263, Dbia5-215, DperSat3-21, and DpseSat2-21 as SatDNA families that did not originate from TEs, we ran BLASTn searches using each SatDNA consensus sequence as queries against *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; TE was not found in regions immediately flanking the SatDNAs. 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 showed several contigs composed exclusively of the 215 bp monomers (e.g., AFFD02001879.1; AFFD02002683.1). The same pattern was observed for the two *D. pseudoobscura* group species (AAIZ01026166.1; AAIZ01022678.1; AADE01010412.1). 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 TE fragments and are incorrectly annotated as TE elements. Altogether, our data suggest SatDNA originating from TE may be particularly common in *D. virilis* group species but are mostly absent in the other species examined.

Evolutionary Maintenance of SatDNA Families Supports the SatDNA Library Hypothesis

The SatDNA library hypothesis posits that the genomes of related species would contain similar families of SatDNA (Fry and Salser 1977). This hypothesis also assumes that multiple SatDNA families can coexist within the same genome, forming a collection of repetitive sequences shared among lineages. Although similar SatDNA may be shared between species, the abundance of SatDNA is likely to stochastically change through both expansion and shrinkage in closely related species (Mestrović et al. 1998). A clear SatDNA library landscape is observed in *D. pseudoobscura* and *D. persimilis* genomes which share four SatDNA families; strikingly, the abundance of each SatDNA family varied significantly between the two species, demonstrating the rapid changes in SatDNA content in the genomes of closely related species (fig. 5*A*). Two of the four SatDNA families characterized in *D. pseudoobscura* and *D. persimilis* (21 and 319 bp) are also present in the *D. miranda* genome (Mahajan et al. 2018), indicating even longer maintenance of both SatDNA families in this species group. A SatDNA library pattern is also observed in the *virilis* group, where we confirmed that the SatDNA family *pvB370* is conserved for a period of about 20 Myr in the *D. virilis* group species with different abundances in each species (Heikkinen et al., 1995; Biessmann et al. 2000). We also confirmed the existence of a 172 bp SatDNA family previously described by Abdurashitov et al. (2013) and Silva et al. (2019). Additionally, we described one new SatDNA family with 393 bp shared among the four species of the *D. virilis* group (supplementary table S2, Supplementary Material online). As expected from the SatDNA library hypothesis, all four SatDNA families comprise different genomic proportions in all four species that diverged ∼10 Ma, demonstrating that nonhomologous SatDNA families can be shared and independently amplify and contract in each genome (fig. 5*B*).

A remarkable example of a SatDNA library pattern was observed between *D. albomicans* and *D. nasuta*. Both species share a recent common ancestor (0.5 Ma) but *D. albomicans* has a very recent neo-Y chromosome system (∼0.1 Ma) (Bachtrog 2006; Wei and Bachtrog 2019; Mai et al. 2020). To better understand the impact of neosex chromosomes in the SatDNA content, we compared both female and male genomes of *D. albomicans* to *D. nasuta* and observed significant changes in the total amount of SatDNA sequences, although both species and sexes share the same SatDNA families (the exception is the lowest abundant DnasSat7–444, detected only in *D. nasuta* genomes) (fig. 5*E* and *F*). We found that *D. albomicans* and *D. nasuta* male genomes have the highest SatDNA content 39.1% and 33.9%, followed by 18.1% and 16.5% in females, respectively (fig. 5*E* and *F*). The comparative relative abundance of SatDNA between male and female genomes from *D. albomicans* and *D. nasuta* shows that the SatDNA doses differ significantly between sexes for both species, with males harboring a higher copy number than females for most (except for DnasSat2–171 in females, see below) of repeats described.

Finally, Cactophilic *Drosophila* seems to be the species group with the lowest amount of SatDNA (de Lima et al. 2017). Consistent with the SatDNA library hypothesis, the pool of ancestral SatDNA is maintained throughout this phylogenetic group (fig. 5*C*). The identification of the *pBuM* SatDNA family in *D. mojavensis*, *D. arizonae*, and *D. buzzatii* indicates the existence of this SatDNA in the shared common ancestor (12 Ma) of these species (de Lima et al. 2017).

Birth and Expansion of New SatDNA Over Evolutionarily Short Timescales

Despite the maintenance of SatDNA families over prolonged periods, we also observed the independent birth/

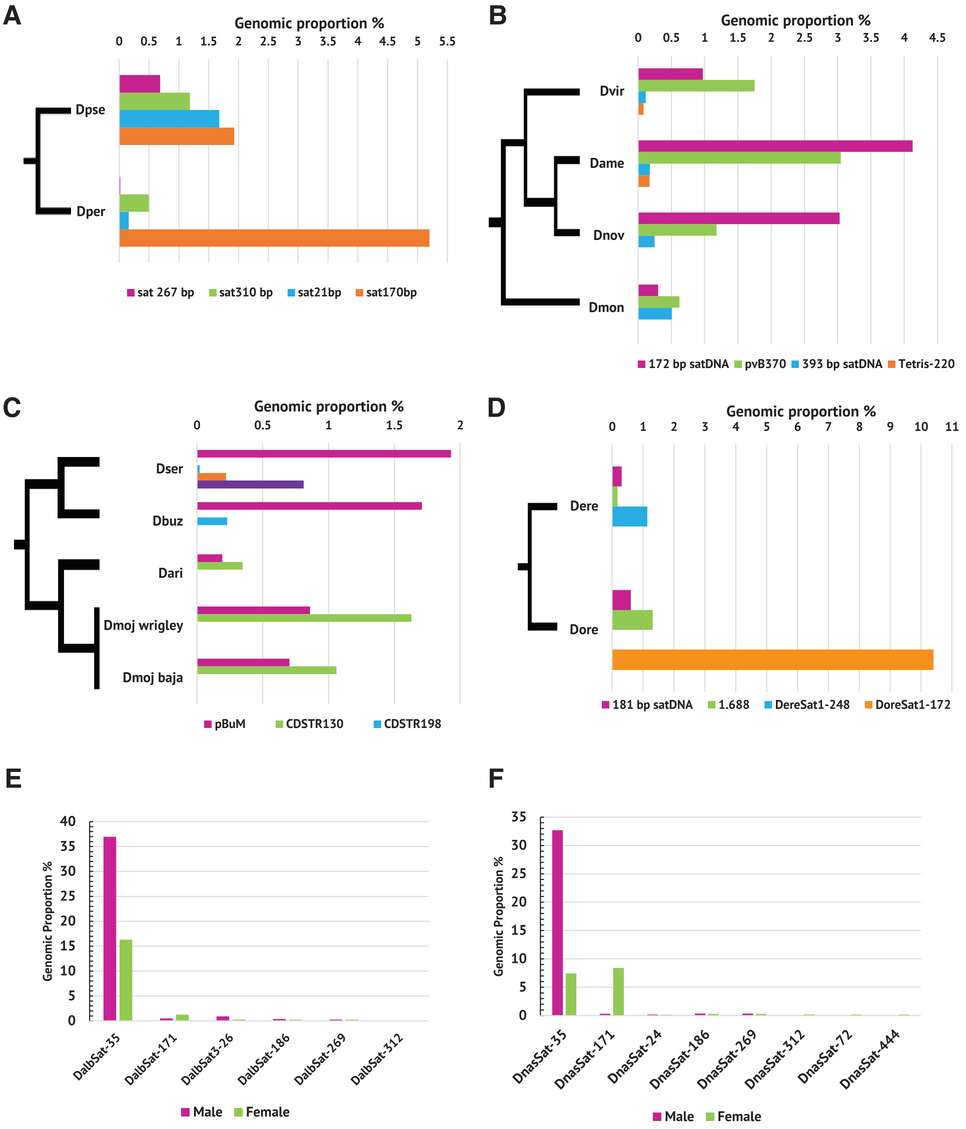


Fig. 5.—Maintenance and genomic proportions of shared SatDNA families among closely related species support the SatDNA library hypothesis. We show the variation in SatDNA library profiles among closely related species from (*A*). *D. pseudoobscura* subgroup; (*B*) *D. virilis* subgroup; (*C*) *D. repleta* subgroup; and (*D*) *D. orena-D. erecta* species. (*E* and *F*) SatDNA families profiles and genomic proportion between male and female specimens of *D. albomicans* and *D. nasuta*, respectively.

expansion of new families and turnover events that have changed the abundance of these sequences over evolutionarily short timescales in *Drosophila* species. For instance, the closely related species *D. orena*, *D. erecta*, *D. yakuba*, *D. teissieri*, and *D. santomea* share a common ancestor ~5.7 Ma (David et al. 2007), yet *D. orena* has the largest genome in the *D. melanogaster* subgroup (280.7 Mb), almost 100 Mb larger than the *D. melanogaster* genome (fig. 2). Interestingly, *D. orena* shows 6–10-fold more SatDNA content than the other four species. We found that the *D. orena* genome has a SatDNA family (DoreSat1–172) that represents 10.4% of its genome. DoreSat1–172 sequences can be found in low copy numbers in *D. erecta* and *D. yakuba* genomes (<0.01% of each species genome and therefore not fully characterized in this study), indicating that this sequence was present in the common ancestor of these species and passed through a recent expansion only in *D. orena* (fig. 5*D*). Similar patterns of independent SatDNA family birth/expansion were observed between closely related species such as the already cited *simulans* clade, *D. buzzatii*/*D. seriema*, *D.albomicans*/*D. nasuta*, and in the *D. yakuba* species complex (*D. yakuba*, *D. teissieri*, and *D. santomea*), in which each species has species-specific SatDNA families, that is,

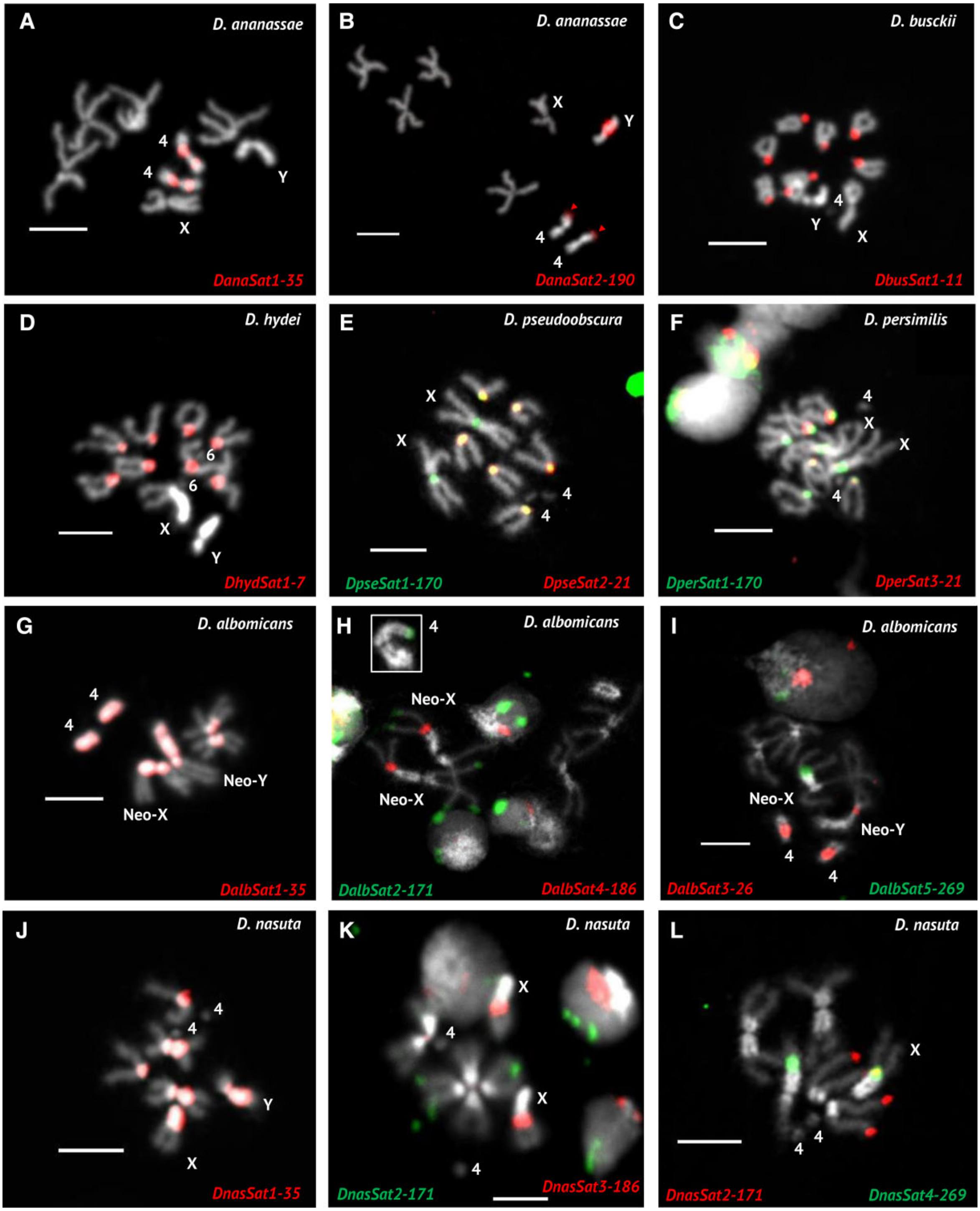


Fig. 6.—Chromosomal localization and distribution of 16 novel SatDNA repeats on mitotic chromosomes of seven *Drosophila* species. FISH was performed on neuroblast chromosome spreads from *D. ananassae*, *D. busckii*, *D. hydei*, *D. pseudoobscura*, *D. persimilis*, *D. albomicans*, and *D. nasuta*. (*A* and *B*) SatDNA DanaSat1–35 and DanaSat2–190 in *D. ananassae*, respectively; (*C*) DbusSat1_11 in *D. busckii*; (*D*) DhydSat1_7 in *D. hydei*; (*E*) DpseSat1_170 and DpseSat2_21 in *D. pseudoobscura*; (*F*) DperSat1_170 and DperSat3_21 in *D. persimilis*; (*G–I*) *DalbSat1–35*, *DalbSat2–171*, *DalbSat3–26*, *DalbSat4–186*, and *DalbSat5-269* in *D. albomicans*; (*J–L*) *D. nasuta: DnasSat1–35*, *DnasSat2–171*, *DnasSat3–186*, and *DnasSat4–269*. Chromatin was stained with DAPI and artificially colored gray, oligo-DNA probes to each SatDNA are labeled as described in supplementary table S4, Supplementary Material online. The sex chromosomes (X and Y) and dot chromosome pairs (4 or 6) are identified in each panel. Bar = 5 μm. Red arrowheads in (*B*) indicate dim signals of hybridization present at telomeric/subtelomeric regions of dot chromosomes pair (4) in *D. ananassae*.

not shared between closely related species (supplementary table S2, Supplementary Material online). One notable example of the birth of a new SatDNA sequence is observed for the 193–7 bp SatDNA shared among *simulans* clade species (Chakraborty et al. 2021). Blast searches in 60 assembled genomes of *D. melanogaster* using all 193–7 bp consensus as query indicate the presence of one single copy sequence of 192 bp present in a syntenic position on the X chromosome (supplementary material S5, Supplementary Material online). These results suggest that one single copy sequence present on the X chromosome of the *simulans* clade species common ancestor passed through a recent expansion event. Overall, our results reinforce the idea that SatDNA family births and amplifications are a common feature of eukaryotic genomes and contribute, concomitantly with the SatDNA library, to the dynamics of the genomic landscape of *Drosophila* species.

Chromosomal Mapping and Validation of 16 Newly Identified SatDNA Throughout *Drosophila*

To confirm and extend our computational analysis, we experimentally mapped the distribution of 17 newly identified SatDNA throughout *Drosophila* species by performing fluorescent in situ hybridization (FISH) experiments using species-specific probes (supplementary material, Supplementary Material online) on the mitotic chromosomes of squashed larval brains from seven *Drosophila* species: *D. albomicans*, *D. ananassae*, *D. busckii*, *D. hydei*, *D. nasuta*, *D. persimilis*, and *D. pseudoobscura* (fig. 6). This set of species was selected aiming to broadly cover the main branches of the genus' phylogeny (fig. 1) and fill the gaps of previous satDNA mapping cytogenetics analyses. The most abundant SatDNA described for each species showed hybridization signals in the pericentromeric regions of multiple chromosomes (fig. 6), except for DanaSat1–35 which is exclusive to the enlarged dot chromosomes pair in *D. ananassae* (fig. 6*A*).

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*		*P*-Value
SatDNA content/ genome size	*Drosophila* + *Sophophora*	34	0.677389778	3.60E−05
	Drosophila	11	0.836827409	0.000627448
	Sophophora	23	0.334782609	0.109803536
Repetitive content/ genome size	*Drosophila* + *Sophophora*	34	0.66748167	3.16E−05
	Drosophila	11	0.786617764	0.000627448
	Sophophora	23	0.571304348	0.003544699
SatDNA content/ repetitive content	*Drosophila* + *Sophophora*	38	0.568882218	0.001158157
	Drosophila	14	0.961538462	0.000193554
	Sophophora	24	0.247863248	0.222137589

Aiming to confirm the presence of both high and low abundant SatDNA families, we conducted FISH experiments in *D. albomicans* and *D. nasuta* mitotic metaphases for SatDNA families with predicted genomic proportion varying from 0.22% to 36.6% (figs. 5E, F and 6; supplementary material, Supplementary Material online). DalbSat1–35 SatDNA family is present in all three chromosome pairs and shows a remarkable presence in the neo-X and the neo-Y chromosomes. The abundance concerning the number of loci on the neosex chromosome was also notable when compared with autosomes and correlates with the genomic proportion described for both males and females. Among the five SatDNA families mapped in *D. albomicans*, four of them (DalbSat1–35, DalbSat2–171, DalbSat4–186, and DalbSat5–269) are found in the neo-X chromosome, whereas DalbSat3–26 is present at ChrY and Chr4 loci (fig. 6; supplementary material, Supplementary Material online). Interestingly, DalbSat4–186 and DalbSat5–269 arrays are present in close distal portions of heterochromatin, whereas DalbSat2–171 signals are observed in the telomeric/subtelomeric regions of Chr3 and neo-X chromosome. Despite the different karyotypic organizations, a similar pattern of chromosomal localization (Muller elements) is observed in *D. nasuta* for DnasSat1–35, DnasSat3–186, and DnasSat5–269. The highly abundant DnasSat1–35 (up to 32.7% in males) is present in all pericentromeric regions, except the dot chromosomes, whereas DnasSat3–186 and DnasSat5–269 are restricted to the X chromosome distal heterochromatin. We identified DnasSat2–171 signals at telomeric/subtelomeric regions of chromosome 3; however, we could not detect signals for the predicted arrays of DnasSat2–171 present at X or 2R chromosomes. Our phylogenetic analysis using a total of 910 monomers for the DalbSat2–171 and DnasSat2–171 SatDNA families indicated that the chromosome X arrays are significantly different from those arrays in other chromosomes (supplementary material S5, Supplementary Material online). FISH experiments hybridizing chromosome-specific probes will be required to confirm the localization of DnasSat2–171 at X or 2R chromosomes. Our cytogenetic results corroborate the dramatic variation in abundance predicted in silico and validate the characterization of eight simple and nine complex SatDNA families in seven different species.

The Genome Size of *Sophophora* and *Drosophila* Subgenera Correlates with Repetitive DNA Content

Given the ~2.5-fold variation in genome size observed in the *Drosophila* species analyzed, we expected to find evidence of repetitive elements substantially contributing to nuclear DNA content (Gregory and Johnston 2008; Dodsworth et al. 2014). The SatDNA abundance and the overall repetitive DNA content for 37 species are shown

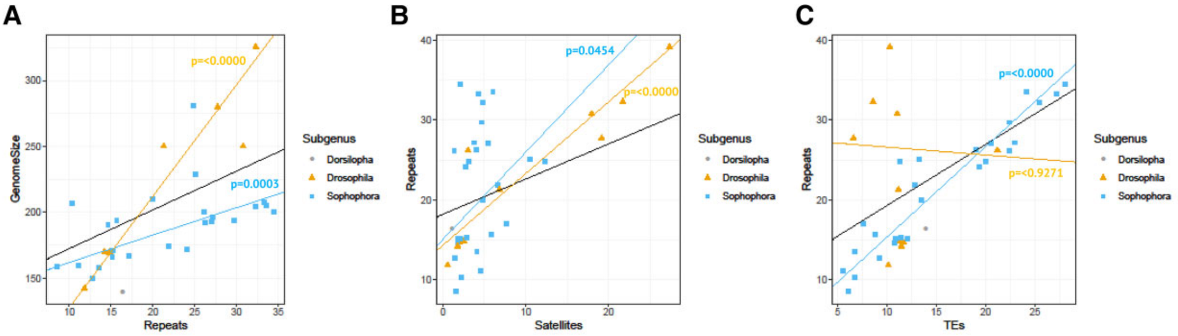


Fig. 7.—PGLS correlation of SatDNA and TEs with genome size in *Drosophila* genus and both major subgenera *Drosophila* and *Sophophora* suggest a differential influence of repetitive components in the genome size variation. PGLS linear regression plots and coefficient between (*A*) repetitive DNA content and genome size, (*B*) SatDNA content and repetitive content, and (*C*) TE content and repetitive Content variation in the *Dorsilopha* subgenus, and both subgenus *Sophophora* and *Drosophila* independently. Regression plots show a significant association of SatDNA variation and genome size evolution in subgenus *Drosophila*, whereas non-SatDNA repeats (mostly TE) strongly correlate with genome size variation in *Sophophora*.

in table 2. The genome size of the *Drosophila* species analyzed ranges from 139.9 to 325.4 Mb, with an average of 196.2 Mb and a median of 193 Mb. The variation in SatDNA content and its contribution to genome size are consistent with previous ideas that genomes have expanded/contracted mainly by the addition/deletion of repeated sequences (for a review see Gregory 2005). Our estimates suggest that SatDNA sequences comprise a lower proportion of the genome compared with previous analyses in *Drosophila* (Bosco et al. 2007; Craddock et al. 2016).

Prompted by the results described above, we ran statistical analyses using phylogenetic generalized least squares (PGLS) (Revell 2010) to systematically investigate the correlation between SatDNA content, repetitive DNA, and genome size (fig. 7). We investigated whether the genome size of each subgenus is equally well correlated with SatDNA and overall repetitive content. The positive relationship between genome size and SatDNA is strongly significant in the *Drosophila* subgenus even after correcting for the phylogenetic groupings. A similar coefficient was observed for SatDNA and repetitive DNA content and the correlation between repetitive DNA content and genome size in the *Drosophila* subgenus. In contrast, the *Sophophora* species SatDNA content is less positively correlated with its genome size. The correlation between repetitive DNA and SatDNA content is also low, suggesting that other repetitive sequences, such as TEs, are the major components of genomes in the *Sophophora* subgenus (supplementary material, Supplementary Material online). The phylogenetic signal is confirmed by generalized least squares analysis for all three datasets and the Brownian model (akaike information criterion [AIC] = 147.02) better fits the genome size evolution association with repetitive DNA and SatDNA than the nonphylogenetic (AIC = 345.5161) or the constrained evolution Ornstein–Uhlenbeck (AIC = 328) model. Our results suggest that genome size evolution in *Drosophila* can be regulated by both SatDNA and TEs, and the predominant repetitive DNA may depend on the evolutionary backdrop.

Discussion

In the present study, we have performed the largest satellitome identification and characterization in *Drosophila* species to date. In summary, the de novo characterization of 37 *Drosophila* satellitomes using the *RepeatExplorer* pipeline has confirmed the presence of previously identified sequences and described 112 new SatDNAs, significantly broadening the knowledge of SatDNA sequences throughout *Drosophila* phylogeny. Our results indicate that the *Drosophila* SatDNA landscape is variable in terms of monomeric size, nucleotide composition, and overall genomic proportion. Notably, most species' genomes contained more than one SatDNA family and many SatDNA families show a low genome proportion, as generally observed in insects (Palomeque and Lorite 2008; Ruiz-Ruano et al. 2018; Palacios-Gimenez et al. 2020). Moreover, we described a trimodal distribution of SatDNA monomer sizes in which the simple SatDNA families (<50 bp) are more frequent than complex SatDNA (fig. 3*A*), suggesting that a significant subset of SatDNA monomers is not compatible with the nucleosome wrapping hypothesis (Henikoff et al. 2001). Different features such as secondary or tertiary structures may influence the preferential motif size of these SatDNA families (Garavís et al. 2015; Patchigolla and Mellone 2021). Our results also indicate that simple SatDNA families can comprise a large fraction of a genome (e.g., *D. virilis* clade, *D. hydei*, *D. albomicans*, and *D. nasuta*). Computational approaches based on k-mer searches might better describe simple SatDNA families/subfamilies with high sequence similarity (>80%) within a genome as observed in *D. melanogaster* (see Results;

supplementary table S3, Supplementary Material online) (Wei et al. 2014, 2018).

Our analyses suggest that SatDNA originating from TE structural repeats are particularly common in the *D. virilis* group species but are mostly absent in the other species analyzed. We confirmed the presence of three SatDNA families that originated from TEs in the *D. virilis* group (*pvB370*, *Tetris-220*, and *DINE-TR1 CTRs*). This finding suggests that amplifications from extant abundant sequences could play a key role in the evolution of the genomic landscape of the *D. virilis* group for reasons still to be elucidated. Moreover, the "hybrid" structure of mobile elements incorporating tandem repeats in *Drosophila*, such as Helitrons (Dias et al. 2015), results in complications in the precise classification and quantification of SatDNA sequences (see Results, Šatović Vukšić and Plohl 2021). Therefore, meticulous sequence characterization of *RepeatExplorer* outputs is required to surpass this significant classification challenge (Montiel et al. 2021; Cintra et al. 2022).

In this study, we report that SatDNA sequences constitute a variable portion of each *Drosophila* genome. Genome size and repetitive DNA content covaried across *Drosophila* phylogeny (Hjelmen and Johnston 2017). SatDNA content and the SatDNA families comprising it were shared among closely related species (figs. 2 and 5). These findings support the SatDNA library hypothesis and show that different SatDNA families coexist within the same genome as part of a collection of repetitive sequences shared among closely related species (Fry and Salser 1977) (fig. 5). We also discovered multiple independent births and/or expansion events of SatDNA families over evolutionarily short timescales (e.g., *D. yakuba*; *simulans* clade; *D. buzzatti*/*D. seriema*; *D. albomicans*/*D. nasuta*). We acknowledge that the definition of the seminal sequence and amplification history of a SatDNA family represents a challenge to the field and increases the difficulty to test the SatDNA library hypothesis. However, the characterization of SatDNA families that are absent (to the current level of resolution) or lack significant genomic representation (<0.01%) in sister species suggests that the birth or amplification, respectively, of new SatDNA families can occur even throughout short evolutionary timescales in *Drosophila.* We indicate that the birth of new satDNA families during recent cladogenetic events (as demonstrated in *D. simulans* clade) can also play a role in the repetitive DNA landscape of a group of species, resulting in an addendum to the satDNA library hypothesis.

Generating the SatDNA content of each species enabled us to analyze the correlation between SatDNA and genome size in *Drosophila* species. The genome size variation in *Drosophila* species is postulated to be associated with the amplification/contraction of heterochromatic blocks (Gregory and Johnston 2008; Craddock et al. 2016), comprised mainly of SatDNA arrays and TEs. Our data demonstrate that both SatDNA and non-SatDNA (mostly TE—supplementary material S4, Supplementary Material online) are correlated with genome size (fig. 2) in line with previous proposals that larger genomes have more repetitive content (Bosco et al. 2007). Here, we described that SatDNA content is highly variable in *Drosophila* genomes, and this variation roughly follows phylogenetic relationships in which closely related species tend to show a similar SatDNA content. We observed that genome size variation in the *Drosophila* subgenus shows a strong correlation with the expansion/shrinkage of SatDNA sequences, supporting the hypothesis that SatDNA dynamic changes play a role in genome size evolution (Flynn et al. 2021; Majid and Yuan 2021). Conversely, we described that genome size variation in *Sophophora* correlates with TE content (fig. 7; Sessegolo et al. 2016) and that the sum of all repetitive DNA identified in *Sophophora* does not correlate with the SatDNA abundance (table 2), notably illustrated in *D. affinis* and *D. takahashii* genomes (fig. 2). This particular result suggests that non-SatDNA repetitive sequences, especially TE (fig. 7; supplementary material S4, Supplementary Material online), constitute a major genomic driver for the genome size variation observed in *Sophophora*. This observation is consistent with the published finding that chromosome 4 in *D. ananassae* (subgenus *Sophophora*) is enlarged due to TE insertions (Leung et al. 2017); interestingly, the most abundant SatDNA in this species is also restricted to the chromosome 4 pair (fig. 6*A*). Different expansion/contraction events, such as indels or intron lengths, are also linked to genome size variation in *Drosophila* (Gregory 2003; Kelly et al. 2015), however, our results suggest that genome size evolution in the *Drosophila* subgenus is driven by the modulation of SatDNA sequences, whereas TE is a more prevalent driver in *Sophophora*. The factors driving this difference in these species will be an interesting focus of future investigations, which may include genomic and ecological factors, silencing mechanisms, and host defense systems (Luo and Lu 2017; Funikov et al. 2018; Flynn et al. 2021). It will be fruitful to unveil whether SatDNA/TE content benefits the organism and if genome size variation associated with SatDNA/TE expansions or contractions is influenced by natural selection (Petrov 2001; Graur et al. 2015).

Altogether, we identified 188 SatDNA sequences in 37 species of *Drosophila* with 112 of them being newly characterized, indicating that SatDNA are underrepresented genomic features that are frequently absent in sequence databases. Despite the advances presented here, our understanding of the full spectrum of SatDNA genomic contribution and organization in *Drosophila* is far from complete. There are three important limitations to our study. First, genome assemblies generated by NGS technology tend to collapse the number of SatDNA monomers inside arrays and produce a significant number of unmapped contigs. Also, NGS PCR-based sequencing approaches may hinder or underestimate the genomic representation of simple SatDNA due to amplification biases (Wei et al. 2018). Further cytogenetics and long-read assembly analyses are still necessary to better understand the organization and distribution of SatDNA sequences. Second, SatDNA sequences are a fast-evolving genomic feature, and studies are limited by the species present in the literature, which may display a limited snapshot of SatDNA sequence evolution in *Drosophila*. Finally, considerable intraspecific SatDNA variation is observed in eukaryotic genomes (Arora et al. 2021; de Lima et al. 2021), thus future population surveys of genome size variation and repeat composition may be necessary to completely understand the evolution of SatDNA in *Drosophila*. Our results help reveal how SatDNA evolved and are evolving within the *Drosophila* genus and serve as an important resource to improve analysis of SatDNA in *Drosophila* genome assemblies, comparative genomic analyses, and future functional studies.

Materials and Methods

Genomic Data and Characterization of SatDNA Families

We used genomic Illumina libraries from the 37 *Drosophila* species publicly available in the EBI Short Read Archive with accession numbers described in supplementary table S1, Supplementary Material Online. Moreover, we retrieved the genome size estimations used in the present study in the Animal Genome Size Database (www.genomesize.com) (Bosco et al. 2007; Hjelmen and Johnston 2017; Hjelmen et al. 2019).

We characterized repetitive elements using the software *RepeatExplorer* version1 implemented in Galaxy (http://www.RepeatExplorer.org/) (Novák et al. 2013), which performs a de novo assembly of repetitive elements using a graph-based method to group reads into discrete clusters based on all-by-all BLAST similarity. Before the *RepeatExplorer* analysis, we trimmed reads from all reads used in this analysis to 100 bp and removed sequencing adapters with Trim Galore! (version 0.6.4_dev; https://github.com/FelixKrueger/TrimGalore/releases/tag/0.6.4) and Cutadapt (Martin 2011) software, respectively, except for *D. mauritiana* and *D. persimilis*, with read lengths of only 76 bp after adapter removal. In addition, we set the cut-off quality value at Q30 and kept only reads with 90% of bases with values higher than this value. Then, as part of the *RepeatExplorer* analysis, we explicitly chose a stricter clustering threshold of 90% of the read length instead of 65% to uniquely characterize each repetitive sequence. We considered in this study only clusters with genomic proportion equal or higher to 0.01% as recommended by *RepeatExplorer* authors (Novák et al. 2013). Next, the reads that shared high sequence similarity were clusterized 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). We ran *RepeatMasker* (Smit et al. 2013–2015) analyses using read sequences within individual clusters against the Repbase Metazoa database (Bao et al. 2015) to provide information for their annotations. 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) similarity searches on the Repbase database.

Moreover, only clusters consisting of at least 20 reads were considered. This cut-off was low enough to fully capture highly abundant repeats in all species yet remained computationally tractable. We manually curated all clusters using visual inspection of the *RepeatExplorer* assembled contigs, BLASTn searches at NCBI nt/nr database, and Tandem Repeats Finder (Benson 1999) with default parameters (except max. motif size 2,000 bp), to identify tandemly organized sequences. We generated dot-plots with the Dotlet applet (Junier and Pagni 2000) with a 15 bp word size and 60% similarity cut-off. Following the tandem repeats identification step, we ran a second round of *RepeatMasker* searches using the *Drosophila* database (https://www.girinst.org/censor/index.php) to characterize possible TE internal repeats. Manual curation was also required to exclude repetitive gene families (e.g., histone genes, rDNA, etc.), mitochondrial and microbial contaminants. To confirm the presence of tandemly repeated arrays in the genome of *Drosophila* species, all SatDNA consensus sequences identified by *RepeatExplorer* methods were used as a query for BLASTn searches on assembled genome data from the NCBI whole genome sequences 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 containing arrays retrieved from BLASTn output with *e*-value lower than 10^{-5}, when available. The nomenclature of all new SatDNA was assigned as suggested by Ruiz-Ruano et al. (2016), which includes: species name abbreviation, number of decreasing abundance order, and the repeat unit size. Thus, we characterized all tandemly arranged sequences in this study after confirming the period size and the nucleotide structure, as well as, discarding other repetitive sequences. This manual annotation was crucial to the reliability of each SatDNA description.

To search for homologous SatDNA between different species, we built a SatDNA database by concatenating all the SatDNA consensus sequences detected in each species and performed an all-against-all comparison of the consensus sequences in the database using the rm_homology.py script (https://github.com/fjruizruano/ngs-protocols). Further, we confirmed independently the repetitive DNA profiles of SatDNA sequences identified by *RepeatExplorer* methods using the *RepeatProfiler* software (github.com/johnssproul/RepeatProfiler) (Negm et al. 2020) with the same set of sequence read archive files (supplementary table S1, Supplementary Material online) with the previous cut-off quality values, trimming, and adapter removal options. We used *RepeatProfiler* with default parameters for generating, visualizing, and comparing repetitive DNA profiles from low-coverage, short-read sequence data. Due to restrictions in the pipeline, all SatDNA consensuses were artificially arranged in tandem to form input sequences equal to or longer than 100 bp.

We searched for SatDNA arrays in the genome assemblies of eight species from NCBI: *D. albomicans* (GCA_009650485.1), *D. ananassae* (GCA_017639315.2), *D. busckii* (GCA_011750605.1), *D. elegans* (GCA_011057505.1), *D. pseudoobscura* (GCA_004329205.1), *D. santomea* (GCA_016746245.2), *D. teissieri* (GCA_016746235.2), *D.virilis* (GCA_007989325.2). Then we run *RepeatMasker* using the genome assembly as a query and the collection of satellites from the own species as a reference.

We classified *RepeatExplorer* clusters with their annotation as Satellites, those used to characterize a SatDNA, multigenic families with homology with a multigenic family gene (see below), and the remaining ones were considered as transposable elements. To characterize clusters of multigenic families, we searched for homology with those genes with *RepeatMasker* in the contigs of *RepeatExplorer* assemblies with a proportion equal to or higher than 0.01% as we did previously. As a reference, we used sequences of *D. melanogaster* downloaded from GenBank. We used the genes of ribosomal DNAs (accession numbers M21017.1 and NR_001793 for the 45S and 5S rDNAs), U snRNAs (accession numbers NR_001599, NR_002513, NR_001670, NR_001933, and NR_002081 for U1, U2, U4, U5, and U6, respectively) and histones (accession number X14215), as well as the whole mitochondrial genome (accession number NC_024511). Because these genes are very conserved, they are valid to get homology based on nucleotides in all the species included in this study. We accepted annotations when the coverage of the contigs matching with the reference covered at least 50% of all read of its *RepeatExplorer* cluster.

Alignment and Phylogenetic Reconstruction

We aligned a pool of nucleotide sequences retrieved from assembled genomes (when available) for each SatDNA family using MUSCLE (Edgar 2004) integrated into the MEGA7 platform (Kumar et al. 2016). We added the previously described sequences from *Drosophila* SatDNA (when available, supplementary material S3, Supplementary Material online) to confirm the sequence previously described initial nucleotides of SatDNA families present on other organisms available in GenBank (supplementary material S3, Supplementary Material online). We determined the nucleotide evolution model to be used in the phylogenetic reconstructions using jModelTest2 (Darriba et al. 2012). This analysis allowed us to determine the evolutionary model T93 as the most common model to best explain our data. Finally, we performed Neighbor-Joining phylogenetic reconstructions using MEGA 7 (Kumar et al. 2016) with 1,000 bootstrap replicates and applied the T93 evolutionary model.

Fly Stocks

All fly stocks were raised on standard Bloomington medium at 25 °C, and male and female third-instar wandering larvae were used. The following fly stocks were obtained from National Drosophila Species Stock Center: *D. albomicans* (#15112-1751.00), *D. ananassae* (#14024-0371.13), *D. hydei* (#15085-1641.04), *D. nasuta* (#15112-1781.01), *D. persimilis* (#14011-0111.49), and *D. pseudoobscura* (#14011-0121.94). *D. busckii* specimens were collected and identified by Prof. Robert Unckless from the University of Kansas-Lawrence.

Mitotic Chromosomes Preparations and In Situ Hybridization Experiments

Mitotic chromosome preparations and FISH experiments were conducted as described in de Lima et al. (2017). The brains from third-instar larvae were dissected in 0.7% sodium chloride and moved to a fresh 20 µl drop of Colcemid (Roche) in darkness, followed by 0.5% sodium citrate hypotonic treatment for 10 min and then fixated for 2 min in 100 µl of fixative solution (45% acetic acid, 100% ethanol). After fixation, the brains were transferred to a 60 µl drop of 60% acetic acid on heated to 65 °C on a heat block for 5 min or until dry. For FISH, each dried slide was fixed with paraformaldehyde 4% solution for 30 min and then transferred to 100% ethanol. Slides were then incubated in 2× SSC at 65 °C for 30 min and dehydrated in 70% and 96% ethanol for 10 min each. Chromosomes were denatured in 0.07 M NaOH solution for 30 s and immediately incubated in 2× SSC solution for 10 min. Slides were then dehydrated in two consecutive 2 min incubations of 70% and 96% ethanol. For each slide, 20 µl of the FISH solution (50% formamide, 10% dextran sulfate, 2× SSC, 100 ng fluorescence-labeled probe) was applied directly to the dried slide and a clean Parafilm 22 mm × 22 mm coverslip was placed on top. All probes used are described in supplementary table S6, Supplementary Material online. Slides were incubated at 37 °C overnight (16–24 h). After the incubation, slides were washed twice in 2× SSC at

37 °C for 5 min each. Slides were then washed three times in SSCT (4× SSC, 0.1% Tween-20), followed by two washes in 1× PBS; each wash was 5 min. Excess liquid was wiped from around the sample area, and each slide was mounted with 5 µl Vectashield + DAPI and a 22 mm × 22 mm coverslip. Coverslip edges were sealed with nail polish and imaged immediately. Chromosome images were acquired with an Inverted Zeiss LSM 780 confocal microscope. All imaging used a 63× objective for mitotic chromosomes. Stacks of deconvolved images were combined in a *z*-projection showing maximum intensity, cropped to the region of interest, recolored, and adjusted for brightness and contrast in FIJI/ImageJ.

Statistical Analyses

To infer the correlations between each feature analyzed (genome size, SatDNA content, non-SatDNA content, and overall repetitive DNA content), we used the Spearman's correlation coefficient (ρ). This correlation test was used due to the nonparametric nature of genome size and repetitive DNA evolution. Thus, the general form of a null hypothesis for a Spearman correlation is given by *H*0: there is no monotonic association between the two variables. The ρ can be 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 the paired score. Statistical analyses were run using PRISM 9 software. Further, we incorporated phylogenetic information to correlations between repeat abundances and genome size using PGLS with the maximum likelihood method implemented in the nlme R package (Pinheriro et al 2020).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors would like to thank Dr Scott Hawley and Dr Jennifer Gerton for the many discussions, suggestions, and corrections made while writing this paper. F.J.R.-R. was also supported by a postdoctoral fellowship from Sven och Lilly Lawskis fond (Sweden) and a Marie Skłodowska-Curie Individual Fellowship (grant agreement 875732, European Commission).

Data Availability

Original data underlying this manuscript can be accessed from the Stowers Original Data Repository at http://www.stowers.org/research/publications/libpb-1700.

Literature Cited

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450(7167):203–218.

Abdurashitov MA, et al. 2013. Medium-sized tandem repeats represent an abundant component of the Drosophila virilis genome. BMC Genomics 14(1):771.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Arora UP, Charlebois C, Lawal RA, Dumont BL. 2021. Population and subspecies diversity at mouse centromere satellites. BMC Genomics 22(1):279.

Bachmann L, Sperlich D. 1993. Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. Mol Biol Evol. 10(3):647–659.

Bachtrog D. 2006. The speciation history of the *Drosophila nasuta* complex. Genet Res. 88(1):13–26.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1):1–6.

Barnes SR, Webb DA, 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. Nucl Acids Res. 27(2):573–580.

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

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

Brand CL, Levine MT. 2021. Cross-species incompatibility between a DNA satellite and a chromatin protein poisons germline genome integrity. bioRxiv.

Brutlag D, Appels R, Dennis ES, Peacock WJ. 1977. Highly repeated DNA in *Drosophila* melanogaster. J Mol Biol. 112(1):31–47.

Bueno GDP, et al. 2021. Cytogenetic characterization and mapping of the repetitive DNAs in *Cycloramphus bolitoglossus* (Werner, 1897): more clues for the chromosome evolution in the genus *Cycloramphus* (Anura, Cycloramphidae). PLoS One 16(1):e0245128.

Burgtorf C, Bünemann H. 1994. Representative and efficient cloning of satellite DNAs based on PFGE pre-fractionation of restriction digests of genomic DNA. J Biochem Biophys Meth 28(4):301–312.

Cabral-de-Mello DC, Zrzavá M, Kubíčková S, Rendón P, Marec F. 2021. The role of satellite DNAs in genome architecture and sex chromosome evolution in Crambidae moths. Front Genet. 12:661417.

Cattani MV, Presgraves DC. 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.

Chakraborty M, et al. 2021. Evolution of genome structure in the *Drosophila simulans* species complex. Genome Res. 31(3):380–396.

Chang C-H, et al. 2019. Islands of retroelements are major components of *Drosophila* centromeres. PLoS Biol. 17(5):e3000241.

Chang CH, Larracuente AM. 2017. Genomic changes following the reversal of a Y chromosome to an autosome in *Drosophila pseudoobscura*. Evolution 71(5):1285–1296.

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

Cintra LA, et al. 2022. An 82 bp tandem repeat family typical of 3' noncoding end of Gypsy/TAT LTR retrotransposons is conserved in *Coffea* spp. pericentromeres. Genome 65(3):137–151.

Craddock EM, Gall JG, Jonas M. 2016. Hawaiian *Drosophila* genomes: size variation and evolutionary expansions. Genetica 144(1):107–124.

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

David JR, Lemeunier F, Tsacas L, Yassin A. 2007. The historical discovery of the nine species in the Drosophila melanogaster species subgroup. Genetics 177(4):1969–1973.

de Lima LG, et al. 2021. PCR amplicons identify widespread copy number variation in human centromeric arrays and instability in cancer. Cell Genomics 1(3):100064.

de Lima LG, Hanlon SL, Gerton JL. 2020. Origins and evolutionary patterns of the 1.688 satellite DNA family in *Drosophila* phylogeny. G3: Genes Genomes Genet. 10(11):4129–4146.

de Lima LG, Svartman M, Kuhn GCS. 2017. Dissecting the satellite DNA landscape in three cactophilic *Drosophila* sequenced genomes. G3: Genes Genomes Genet. 7(8):2831–2843.

Dernburg AF, Sedat JW, Hawley RS. 1996. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. Cell 86(1): 135–146.

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

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

Dodsworth S, et al. 2014. Genomic repeat abundances contain phylogenetic signal. Syst Biol. 64(1):112–126.

Dover GA, Tautz D. 1986. Conservation and divergence in multigene families: alternatives to selection and drift. Philos Trans R Soc Lond B: Biol Sci 312(1154):275–289.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 32(5):1792–1797.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. PLoS Biol. 7(10):e1000234.

Ferree PM, Prasad S. 2012. How can satellite DNA divergence cause reproductive isolation? Let us count the chromosomal ways. Genet Res Int. 2012:430136.

Flynn JM, Hu KB, Clark AG. 2021. Multiple recent sex chromosome fusions in *Drosophila virilis* associated with elevated satellite DNA abundance. bioRxiv.

Flynn JM, Long M, Wing RA, Clark AG. 2020. Evolutionary dynamics of abundant 7-bp satellites in the genome of *Drosophila virilis*. Mol Biol Evol. 37(5):1362–1375.

Fry K, 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.

Fu J, et al. 2019. Identification and characterization of abundant repetitive sequences in *Allium cepa*. Sci Rep. 9(1):1–7.

Funikov SY, et al. 2018. Spontaneous gain of susceptibility suggests a novel mechanism of resistance to hybrid dysgenesis in *Drosophila virilis*. PLoS Genet. 14(5):e1007400.

Gall JG, Atherton DD. 1974. Satellite DNA sequences in *Drosophila* virilis. J Mol Biol. 85(4):633–664.

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

Garavís M, et al. 2015. The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. Sci Rep. 5(1):13307.

Graur D, Zheng Y, Azevedo RB. 2015. An evolutionary classification of genomic function. Genome Biol Evol 7(3):642–645.

Gregory TR. 2003. Is small indel bias a determinant of genome size?. Trends Genet. 19(9):485–488.

Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. Nat Rev Genet. 6(9):699–708.

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

Guillén Y, et al. 2015. Genomics of ecological adaptation in cactophilic *Drosophila*. Genome Biol Evol. 7(1):349–366.

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

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

Hjelmen CE, Blackmon H, Holmes VR, Burrus CG, Johnston JS. 2019. Genome size evolution differs between *Drosophila* subgenera with striking differences in male and female genome size in *Sophophora*. G3: Genes Genomes Genet. 9(10):3167–3179.

Hjelmen CE, Johnston JS. 2017. The mode and tempo of genome size evolution in the subgenus *Sophophora*. PLoS One 12(3):e0173505.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9(9):868–877.

Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. 2017. Comparative analysis of satellite DNA in the *Drosophila* melanogaster species complex. G3: Genes Genomes Genet. 7(2):693–704.

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

Kelly LJ, et al. 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol. 208(2):596–607.

Khost DE, Eickbush DG, Larracuente AM. 2017. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res. 27(5):709–721.

Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 115(30):E7109–E7118.

Kim BY, et al. 2021. Highly contiguous assemblies of 101 drosophilid genomes. Elife 10:e66405.

Kuhn GCS. 2015. Satellite DNA transcripts have diverse biological roles in *Drosophila*. Heredity 115(1):1–2.

Kuhn GCS, Bollgönn S, Sperlich D, Bachmann L. 1999. Characterization of a species-specific satellite DNA of *Drosophila* buzzatii. J Zool Syst Evol Res. 37(2):109–112.

Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 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 Res. 16(2):307–324.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870–1874.

Kunyavskaya O, Dvorkina T, Bzikadze AV, Alexandrov I, Pevzner PA. 2021. HORmon: automated annotation of human centromeres. bioRxiv.

Laird CD, McCarthy BJ. 1968. Magnitude of interspecific nucleotide sequence variability in *Drosophila*. Genetics 60(2):303–322.

Larracuente AM. 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 Evol Biol. 14(1):233.

Lauria Sneideman MP, Meller VH. 2021. Drosophila satellite repeats at the intersection of chromatin, gene regulation and evolution, editors. Satellite DNAs in physiology and evolution. Cham: Springer. p. 1–26.

Leung W, et al. 2017. Retrotransposons are the major contributors to the expansion of the *Drosophila* ananassae Muller F element. G3: Genes Genomes Genet. 7(8):2439–2460.

Lohe AR, Hilliker AJ, Roberts PA. 1993. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. Genetics 134(4):1149–1174.

Luo S, Lu J. 2017. Silencing of transposable elements by piRNAs in drosophila: an evolutionary perspective. Genom Proteomi Bioinform. 15(3):164–176.

Mahajan S, Wei KHC, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young Drosophila Y chromosome using single--molecule sequencing and chromatin conformation capture. PLoS Biol 16(7).

Mai D, Nalley MJ, Bachtrog D. 2020. Patterns of genomic differentiation in the Drosophila nasuta species complex. Mol Biol Evol. 37(1):208–220.

Majid M, Yuan H. 2021. Comparative analysis of transposable elements in genus *Calliptamus* grasshoppers revealed that satellite DNA contributes to genome size variation. Insects 12(9):837.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17(1):10–12.

Mcgurk MP, Barbash DA. 2018. Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. Genome Res 28(5):714–725.

Melters DP, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14(1):R10.

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

Mestrović N, Plohl M, Mravinac B, Ugarković D. 1998. Evolution of satellite DNAs from the genus Palorus—experimental evidence for the "library" hypothesis. Mol Biol Evol. 15(8):1062–1068.

Montiel EE, Panzera F, Palomeque T, Lorite P, Pita S. 2021. Satellitome analysis of rhodnius prolixus, one of the main chagas disease vector species. Int J Mol Sci. 22:6052

Negm S, Greenberg A, Larracuente AM, Sproul JS. 2020. RepeatProfiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles. Mol Ecol Resour. 21(3):969–981.

Nijman IJ, Lenstra JA. 2001. Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. J Mol Evol. 52(4):361–371.

Novák P, et al. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucl Acids Res. 45(12):e111.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinform. 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.

Ohno S. 1972. So much "junk" DNA in our genome. In: Evolution of genetic systems, Brookhaven symp. biol. p. 366–370.

Palacios-Gimenez OM, et al. 2020. Eight million years of satellite DNA evolution in grasshoppers of the genus *Schistocerca* illuminate the ins and outs of the library hypothesis. Genome Biol Evol. 12(3):88–102.

Palomeque T, Lorite P. 2008. Satellite DNA in insects: a review. Heredity 100(6):564–573.

Patchigolla VS, Mellone BG. 2021. Enrichment of non-B-form DNA at *D. melanogaster* centromeres. bioRxiv.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. TRENDS Gen 17(1):23–28.

Picariello O, Feliciello I, Bellinello R, Chinali G. 2002. S1 satellite DNA as a taxonomic marker in brown frogs: molecular evidence that Rana graeca graeca and Rana graeca italica are different species. Genome 45(1):63–70.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. 2020. nlme: linear and nonlinear mixed effects models. R package version 3.1-144. https://CRAN.R-project.org/package=nlme.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA evolution. In: Repetitive DNA. Vol. 7. Karger Publishers. p. 126–152.

Renkawitz R. 1979. Isolation of twelve satellite DNAs from *Drosophila* hydei. Int J Biol Macromol. 1(3):133–136.

Revell LJ. 2010. Phylogenetic signal and linear regression on species data. Methods Ecol Evol. 1(4):319–329.

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

Ruiz-Ruano FJ, et al. 2018. High-throughput analysis of satellite DNA in the grasshopper Pyrgomorpha conica reveals abundance of homologous and heterologous higher-order repeats. Chromosoma 127(3):323–340.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Sci Rep. 6:28333.

Russo CA, Mello B, Frazão A, Voloch CM. 2013. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). Zool J Linnean Soc. 169(4):765–775.

Šatović Vukšić E, Plohl M. 2021. Exploring satellite DNAs: specificities of bivalve mollusks genomes. In: Satellite DNAs in physiology and evolution. Cham: Springer. p. 57–83.

Schmidt T, Heslop-Harrison JS. 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci. 3(5):195–199.

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

Shatskikh AS, Kotov AA, Adashev VE, Bazylev SS, Olenina LV. 2020. Functional significance of satellite DNAs: insights from *Drosophila*. Front Cell Dev Biol. 8:312.

Silva BS, Heringer P, Dias GB, Svartman M, Kuhn GC. 2019. De novo identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines. PLoS One 14(12):e0223466.

Slamovits CH, Cook JA, Lessa EP, 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. Mol Biol Evol. 18(9):1708–1719.

Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. Available from: http://www.RepeatMasker.org/. Accessed January 03, 2021.

Sproul JS, et al. 2020. Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the simulans clade. Mol Biol Evol. 37(8):2241–2256.

Strachan T, Webb D, Dover GA. 1985. Transition stages of molecular drive in multiple-copy DNA families in Drosophila. EMBO J 4(7):1701–1708.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13(1):36–46.

Waring GL, Pollack JC. 1987. Cloning and characterization of a dispersed, multicopy, X chromosome sequence in *Drosophila* melanogaster. Proc Natl Acad Sci U S A. 84(9):2843–2847.

Wei KHC, et al. 2018. Variable rates of simple satellite gains across the *Drosophila phylogeny*. Mol Biol Evol. 35(4):925–941.

Wei KH, Bachtrog D. 2019. Ancestral male recombination in *Drosophila albomicans* produced geographically restricted neo-Y chromosome haplotypes varying in age and onset of decay. PLoS Genet. 15(11):e1008502.

Wei KHC, Grenier JK, Barbash DA, Clark AG. 2014. Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 111(52):18793–18798.

Associate editor: Josefa Gonzalez