genes

MDPI

Article

In Silico Identification and Characterization of Satellite DNAs in 23 *Drosophila* Species from the *Montium* Group

Bráulio S. M. L. Silva, Agnello C. R. Picorelli and Gustavo C. S. Kuhn *

Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil
* Correspondence: gcskuhn@ufmg.br

Abstract: Satellite DNA (satDNA) is a class of tandemly repeated non-protein coding DNA sequences which can be found in abundance in eukaryotic genomes. They can be functional, impact the genomic architecture in many ways, and their rapid evolution has consequences for species diversification. We took advantage of the recent availability of sequenced genomes from 23 *Drosophila* species from the *montium* group to study their satDNA landscape. For this purpose, we used publicly available whole-genome sequencing Illumina reads and the TAREAN (tandem repeat analyzer) pipeline. We provide the characterization of 101 non-homologous satDNA families in this group, 93 of which are described here for the first time. Their repeat units vary in size from 4 bp to 1897 bp, but most satDNAs show repeat units < 100 bp long and, among them, repeats ≤ 10 bp are the most frequent ones. The genomic contribution of the satDNAs ranges from ~1.4% to 21.6%. There is no significant correlation between satDNA content and genome sizes in the 23 species. We also found that at least one satDNA originated from an expansion of the central tandem repeats (CTRs) present inside a Helitron transposon. Finally, some satDNAs may be useful as taxonomic markers for the identification of species or subgroups within the group.

Keywords: satellite DNA; tandem repeats; repetitive DNA; Helitrons; genome evolution; TAREAN; *Drosophila*; *montium* group

Citation: Silva, B.S.M.L.; Picorelli, A.C.R.; Kuhn, G.C.S. In Silico Identification and Characterization of Satellite DNAs in 23 *Drosophila* Species from the *Montium* Group. *Genes* **2023**, *14*, 300. https://doi.org/10.3390/genes14020300

Academic Editors: Manuel A. Garrido-Ramos, Miroslav Plohl and Eva Šatović-Vukšić

Received: 6 December 2022
Revised: 13 January 2023
Accepted: 16 January 2023
Published: 23 January 2023

1. Introduction

Eukaryotic genomes are enriched by a great number and variety of non-protein-coding repetitive DNA elements. The genomic fraction made by these elements varies between species, but it can frequently reach >50% in several animal and plant species. They can be found dispersed along the genome, in forms such as transposable elements (TEs), and/or in tandem organization, as microsatellites, minisatellites and satellite DNAs (satDNAs) [1,2].

Individual satDNA families typically reach more than 10^3 copies in the genome. These copies form large, in some cases Mb-size arrays, that are mainly concentrated in heterochromatin-rich regions of the chromosomes, such as the (peri)centromeric and subtelomeric regions [2–6]. However, occasionally they may also be present along the euchromatin in the form of small arrays (with 1–20 tandem repeats) [7–9]. In contrast, microsatellites and minisatellites are less repetitive (<10^3 copies), and their shorter arrays are in a scattered distribution throughout the genome. Concerning repeat length (i.e., monomer size), microsatellites are usually in the range of few base pairs to <10 bp, minisatellites between 10 and 200 bp, and satellites between 2 bp to > several hundred bp [3,4].

Once considered fully "junk DNA" in the past, it is now recognized that satDNAs (or a fraction of them) may participate in important genomic functions, such as gene regulation and chromatin modulation [10,11], spatial chromosome organization [12–14], and centromeric architecture [15]. Furthermore, satDNAs contribute to the generation of genome size differences among species and may also be related to the origin of chromosome rearrangements [16,17]. SatDNAs evolve rapidly and may also contribute to the

establishment of genetic incompatibilities and reproductive isolation between incipient species [18]. Therefore, there is no doubt today that the study of satDNAs is highly relevant in the context of functional and evolutionary genomics [6,16,19,20].

Species from the genus *Drosophila* have been extensively used as model to address several aspects related to satDNA structure, organization, function, evolution, and impact on speciation (e.g., [11,14,18,21–26]). In the last 10 years, these studies have been fostered by the large number of *Drosophila* species with sequenced genomes available, and the concomitant development of several new bioinformatic tools specifically designed for the identification of satDNAs, such as the TAREAN (Tandem Repeat Analyzer) pipeline [27]. More recently, the genomes of 23 *Drosophila* species from the *montium* group have been sequenced, but no information about their satDNAs has been reported to date [28].

The *montium* group, with 71 Asian and Australasian species and 23 African species, is the largest clade within the subgenus *Sophophora*. Based on the analyses of morphological (male abdominal pigmentation and genitalia) and chorological traits, the group can be subdivided into seven subgroups (*parvula*, *montium*, *punjabiensis*, *serrata*, *kikkawai*, *seguyi*, and *orosa*) whose phylogenetic relationships have been inferred from three nuclear genes and one mitochondrial gene [29] (Figure 1A).

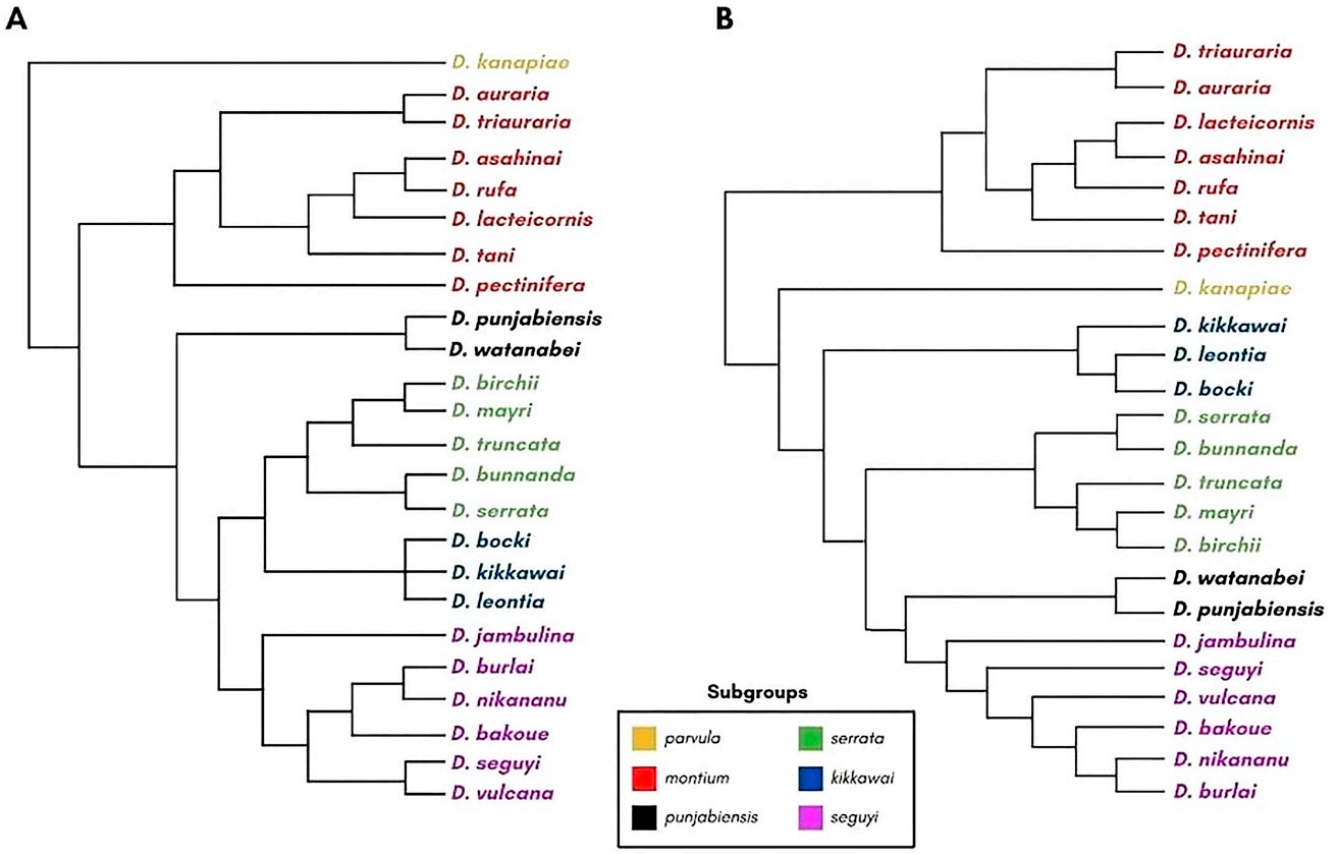

Figure 1. Phylogenetic relationships among subgroups from the *montium* group. Only the species investigated in the present work are shown. (**A**) Phylogeny based on three nuclear genes and one mitochondrial gene (adapted from Yassin [29]). (**B**) Phylogeny based on 60 genes (adapted from Conner et al. [30]). The branch lengths do not correspond to evolutionary distances. The phylogenetic trees were reconstructed using the Archaeopteryx software [31].

A recent phylogenetic analysis, performed using 60 genes made by Conner et al. [30], confirmed the monophyly of the seven *montium* subgroups proposed by Yassin [29]. However, this later study showed that the *montium* subgroup is the most basal subgroup in the phylogeny. Moreover, it showed that the *punjabiensis* subgroup is closer to the

seguyi subgroup and that the *kikkawai* subgroup is the third most basal clade of the group (Figure 1B) [30].

The basic metaphase karyotype of species from the *montium* group consists of one pair of sex chromosomes, two pairs of acrocentric chromosomes and one pair of microchromosomes [32]. The only reported changes in the metaphase karyotype configuration among species concern the variation in the amount of heterochromatin present in the microchromosomes and/or in the Y chromosome and, to a lesser extent, in the X chromosome [32,33]. As found in *Drosophila* and other eukaryotic species, changes in the amount of heterochromatin may be directly connected to expansions or contractions of satDNAs, which is the most abundant component of heterochromatin [34–38].

In the present work, we aimed to characterize the satDNA landscape of 23 species from the *montium* group. We first used the TAREAN pipeline to identify and quantify putative satDNAs in the 23 species, and then created a more conservative "satDNA filter" to select only the families sharing more attributes with satDNAs. We ended up with 142 satDNA clusters representing 101 non-homologous satDNA families. The data are discussed in terms of satDNA's general structural features, its relationship to genome sizes, and its relationship to transposable elements. We expect that our collection of identified satDNAs will be useful for future studies concerning genome annotation and genome/chromosome evolution in the *montium* group. Additionally, some satDNA families may be useful as potential taxonomic markers for the identification of species or specific clades/subgroups within the *montium* species group.

2. Materials and Methods

Satellite DNA Identification

TAREAN is a computational pipeline used for the unsupervised identification of satDNAs from unassembled short-read sequences [27]. In this study, we used publicly available Illumina paired-end sequencing raw data from 23 species (females) from the *montium* group on NCBI (Accession: PRJNA554346—ID: 554346) [28] (Table 1). TAREAN analyses were performed on the Galaxy Platform [39]. We first measured the reads quality with the "FASTQC" tool and converted all the sequences to a single fastqsanger format with the "FASTQ Groomer" tool (Sanger and Illumina 1.8+). After the removal of adapters and reads presenting more than 5% of low-quality bases (Phred cutoff < 10), the reads were trimmed to 100 bp along with the "Preprocessing of fastq paired-reads" tool. The resulting file, with interlaced filtered paired-end reads, was used as an input for the TAREAN pipeline, with the following settings: "read sampling: no—advanced options: yes—perform cluster merging: yes—use custom repeat database: no—cluster size threshold for detailed analysis: 0.01—perform automatic filtering of abundant satellite repeats: no—keep original read names: no—similarity search options: masking of low complexity repeats disabled—select queue: basic". The resulting archives, containing the putative satDNA clusters, were downloaded for a more detailed investigation. Only putative satDNA clusters, showing a minimum of 0.1% genomic contribution to at least one species of the genomic DNA, were selected for further analysis. Considering typical genome sizes of species from the *montium* group as being around 196.3 Mb, 0.1% genomic contribution corresponds to ~1,963,000 copies of satDNA with 10 bp or ~196,300 copies of a satDNA with 100 bp repeats.

The estimated genomic proportion of each putative satDNA cluster is initially presented in the TAREAN results as the proportion of the reads in each cluster concerning the number of all analyzed reads. However, the analyzed reads by TAREAN may contain organellar DNA and contaminant DNA. For this reason, we checked all clusters retrieved by TAREAN in each species and removed (when present) the reads from clusters corresponding to mitochondrial DNA and contaminants. Next, we recalculated the genomic proportion of each putative satDNA based on the number of total reads representing only nuclear sequences, as proposed by Novák et al. [40].

Table 1. SatDNA-like clusters identified in the *montium* group by TAREAN, before and after filtering. HC = High confidence; LC = Low confidence.

Species	Subgroup	HC satDNAs (Before Filtering)	LC satDNAs (Before Filtering)	Final Number of satDNA-like Families (After Filtering)
D. kanapiae	*parvula*	13	9	4
D. auraria	*montium*	3	10	2
D. triauraria	*montium*	6	5	3
D. asahinai	*montium*	7	8	3
D. rufa	*montium*	4	7	3
D. lacteicornis	*montium*	6	5	3
D. tani	*montium*	7	9	4
D. pectinifera	*montium*	14	5	10
D. punjabiensis	*punjabiensis*	14	5	6
D. watanabei	*punjabiensis*	7	7	3
D. birchii	*serrata*	15	5	8
D. mayri	*serrata*	15	8	13
D. truncata	*serrata*	11	6	5
D. bunnanda	*serrata*	24	6	14
D. serrata	*serrata*	12	9	6
D. bocki	*kikkawai*	10	4	7
D. leontia	*kikkawai*	5	7	4
D. jambulina	*seguyi*	8	2	6
D. burlai	*seguyi*	11	7	7
D. nikananu	*seguyi*	6	6	3
D. bakoue	*seguyi*	25	8	9
D. seguyi	*seguyi*	17	6	12
D. vulcana	*seguyi*	5	8	3
Total		245	152	

The TAREAN pipeline classifies the clusters with putative satDNA sequences into two categories: satellites with high confidence (HC) and satellites with low confidence (LC). These categories are determined according to the "Connected component index (C)", which indicates clusters formed by tandem repeat sequences, and "Pair completeness index (P)", which measures the length of continuous tandem arrays [27]. Another important aspect of TAREAN is that the pipeline groups the reads into clusters according to their sequence similarity. Similar reads form graphs represented by nodes and connecting edges, and graphs presenting globular shapes are likely constituted by satDNAs [27].

After selecting the satDNA clusters with more than 0.1% genomic contribution, we developed a second satDNA "filter" in which the selected clusters should comply with three out of the four following parameters: c value > 0.9, p value > 0.8, high confidence and circular graph layout (e.g., Figure S1). After this cut-off analysis, we conducted further analyses on the remaining satDNA clusters and their corresponding consensus sequences provided in the TAREAN results.

For the identification of homologous satDNAs shared by two or more species, we created a custom database on the Geneious Software [41] containing all consensus sequences from the selected satDNAs. We then used each satDNA consensus for MEGABLAST searches of the custom database (maximum e-value = 1×10^{-5}; gap cost = linear; threshold = 0%; majority: most common bases, fewest ambiguities). Figure 2 shows the workflow chart of our study.

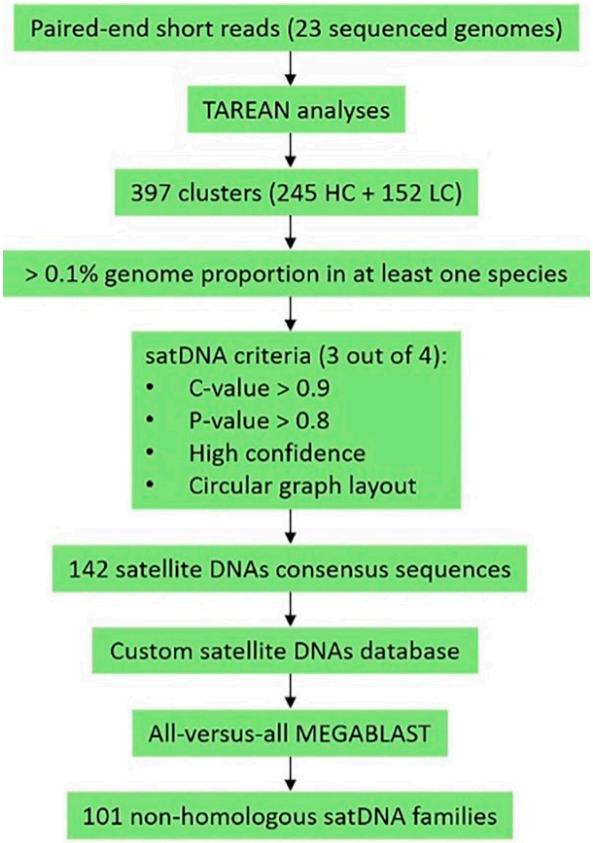

Figure 2. Workflow used for satellite DNA (satDNA) identification in the 23 sequenced *montium* genomes.

3. Results and Discussion

3.1. Identification of Satellite DNA Families in the Montium Group

The TAREAN analysis first retrieved 397 clusters, identified as putative satDNAs in the 23 species from the *montium* group, namely, 245 with high confidence (HC) and 152 with low confidence (LC) (Table 1). After filtering the clusters using our custom satDNA filter, the number of satDNAs narrowed down to 142 (Figure 2). Then, we created a custom database containing consensus sequences of each one of these 142 satDNAs and conducted MEGABLAST searches using each consensus sequence against our whole custom database. We found that the 142 satDNAs correspond to 101 satDNA non-homologous families, which have been numbered dmgsat-1 to dmgsat-101 (Table S1). The consensus sequences of these satDNAs can be found in File S1.

It is assumed that the *c* and *p* values retrieved from TAREAN analyses are important parameters for a reliable satDNA identification, as both values together give a good indication that the identified clusters correspond to repeats, organized as long and continuous satDNA-like arrays. For example, several studies showed that clusters with high genome proportion (>1%), high *c* and *p* values (>0.98) and high satellite probability (>0.95) correspond to typical satDNAs sequences that are located on the centromeric and/or pericentromeric regions of the chromosomes [42–45]. Accordingly, all satDNAs selected for our study have *c* and *p* values near or above 0.9 in at least one species (Table S1).

To our knowledge, from all the 101 satDNA families we found in the *montium* group, only 8 families showed any homology with previously described satDNAs in other *Drosophila* species (Table S1): the dmgsat-14 and dmgsat-67 satDNA families share sequence homology to the "1.688" satellite DNA [7,9,46], and the dmgsat-52 satDNA family is homologous to the "1.669" satDNA [22,47,48]. Recently, de Lima and Ruiz-Ruano [49] reported an in silico characterization of satellite DNAs in two species from the *montium* group,

D. burlai and *D. leontia*, using the RepeatExplorer pipeline. We have noted that five satDNAs reported here are homologous with satDNAs reported in de Lima and Ruiz-Ruano [49]: the dmgsat-10 and dmgsat-11 are homologous to "DleoSat1-41" and "Dleosat4-109" from *D. leontia*, respectively, and dmgsat-51, dmgsat-61, and dmgsat-85 are homologous to "DburSat3-9", "DburSat2-300" and "DburSat1-135" from *D. burlai*, respectively.

3.2. Satellite DNAs in the Montium Group: General Structural Features

There is an extensive variation in repeat length in the 101 satDNAs found in species of the *montium* group, from only 4 bp (dmgsat-35 from *D. triauraria*) to 1897 bp (dmgsat-63 from *D. burlai*) (Figure 3). However, most satDNAs (89%) are within the range of the most common repeat length found in *Drosophila* (from <10 bp to 400 bp) [26,50,51].

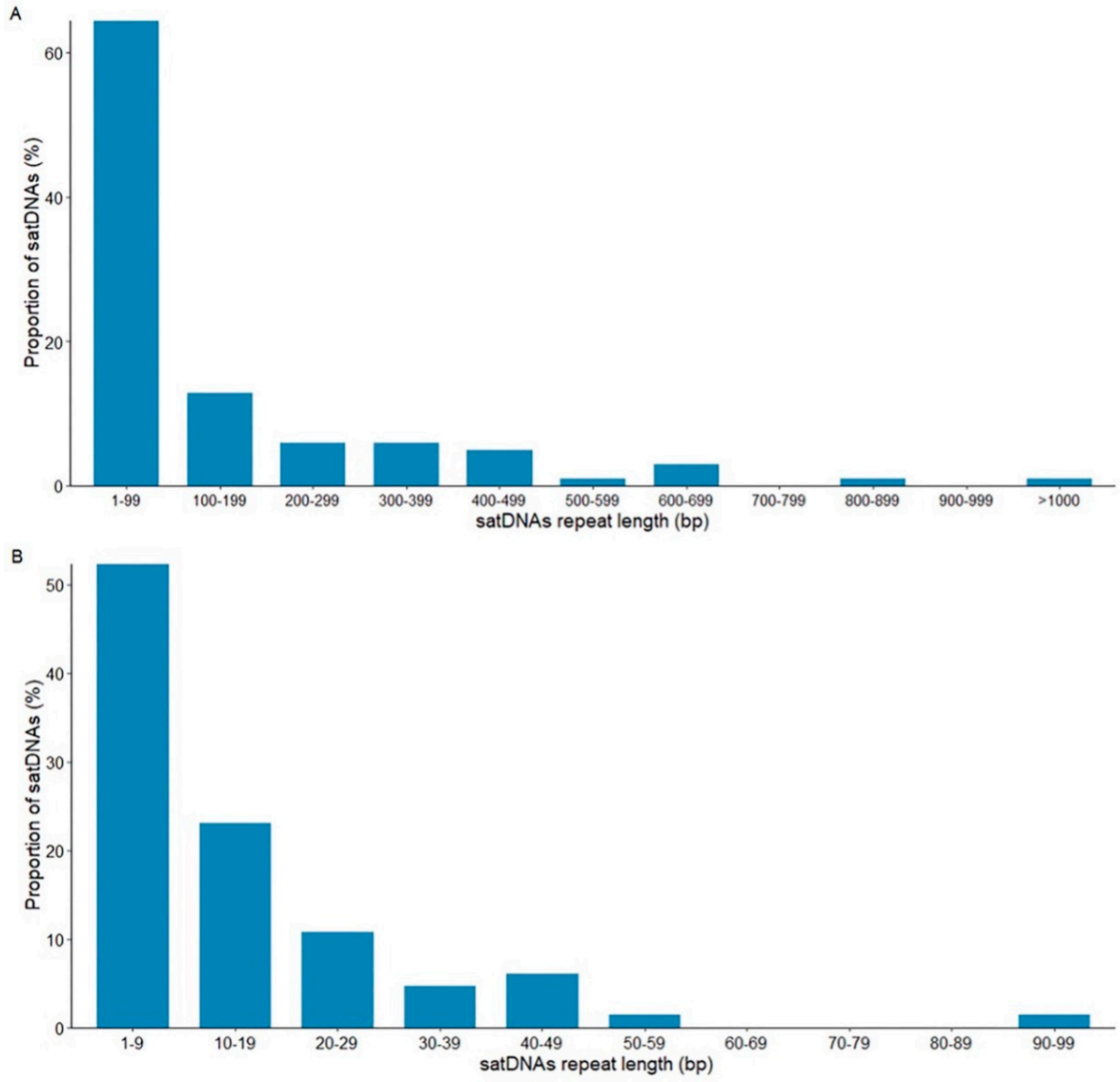

Figure 3. SatDNAs repeat length (monomer size) in the *montium* group. (**A**) The repeat length of all 101 satDNA families identified in the present work. (**B**) Repeat length of the 65 satDNA families featuring less than 100 bp long repeats.

Most satDNAs (60%) showed repeats shorter than 100 bp (Figure 3A). To better assess the repeat length variation of the 65 satDNAs with repeats shorter than 100 bp, we further subdivided this class into 10 intervals of 10 bp each (Figure 3B). Most short satDNA families have repeat sizes shorter than 10 bp (52.3%) (Figure 3B).

Therefore, we concluded that the 23 species genomes from the *montium* group investigated in our study are enriched with satDNAs consisting of short (<100 bp) tandem repeats, especially in the range of 1–9 bp. The presence of satDNAs with short repeat sizes in *Drosophila* is not rare. For example, abundant satDNA families with repeats 7 bp long are found in *Drosophila virilis* [52,53], and *D. melanogaster* has several satDNAs with repeats in the range between 5 bp to 10 bp [22].

We found that 78 satDNA families have an AT content > 60% (Figure 4). This number represents 76.5% of the total number of families. Therefore, our findings show that satDNA sequences, present in species from the *montium* group, are also mostly AT rich, as found previously for other groups and species of *Drosophila* [22,48,49,52–55].

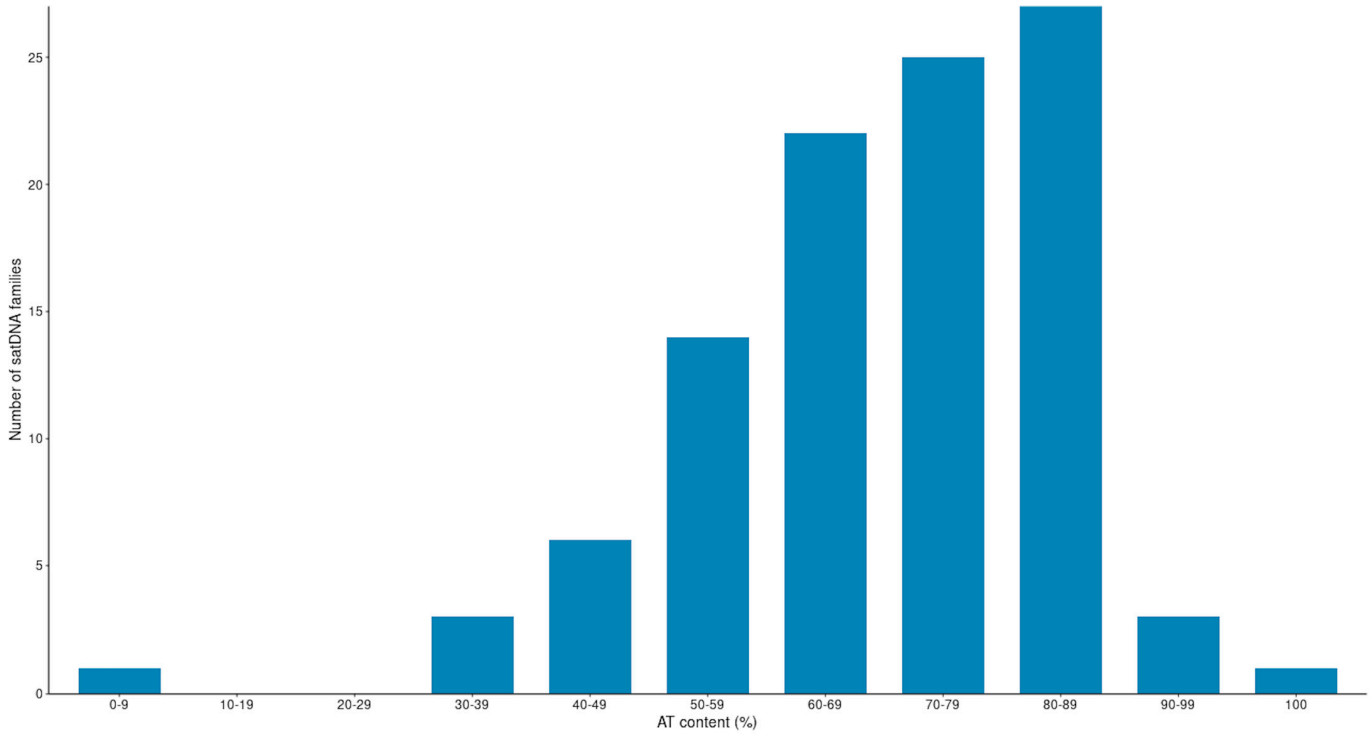

Figure 4. AT content of the 101 satDNA families identified in the *montium* group.

3.3. Satellite DNA Abundance and Relationship with Genome Sizes

SatDNAs usually account for more than 20% of the genomic DNA in species from the *Drosophila* genus [56], as in *D. melanogaster*, and up to 70% in some Hawaiian *Drosophila* [57], but less than 3% in species from the *repleta* group [26]. In *Drosophila* and many organisms, there is a positive correlation between satDNA content and genome sizes [49,56].

The genome sizes in the 23 *Drosophila* studied species from the *montium* group were estimated by Bronski et al. [28] and they range from 155.1 Mb (*D. bocki*) to 223.4 Mb (*D. mayri*). Based on the TAREAN results, our estimated satDNA contribution to total genomic DNA ranges from 1.40% (*D. watanabei*) to 21.65% (*D. pectinifera*) (Figure 5). Such 16-fold variation does not match the 1.4-fold variation found among genome sizes. Accordingly, we found no significant positive correlation between satDNA abundance and genome sizes (Figure 6). We also performed correlation tests between genome sizes and all initial 397 putative satDNA clusters returned by the TAREAN analysis (Figure 2), but again we found no significant correlations (Figure S2).

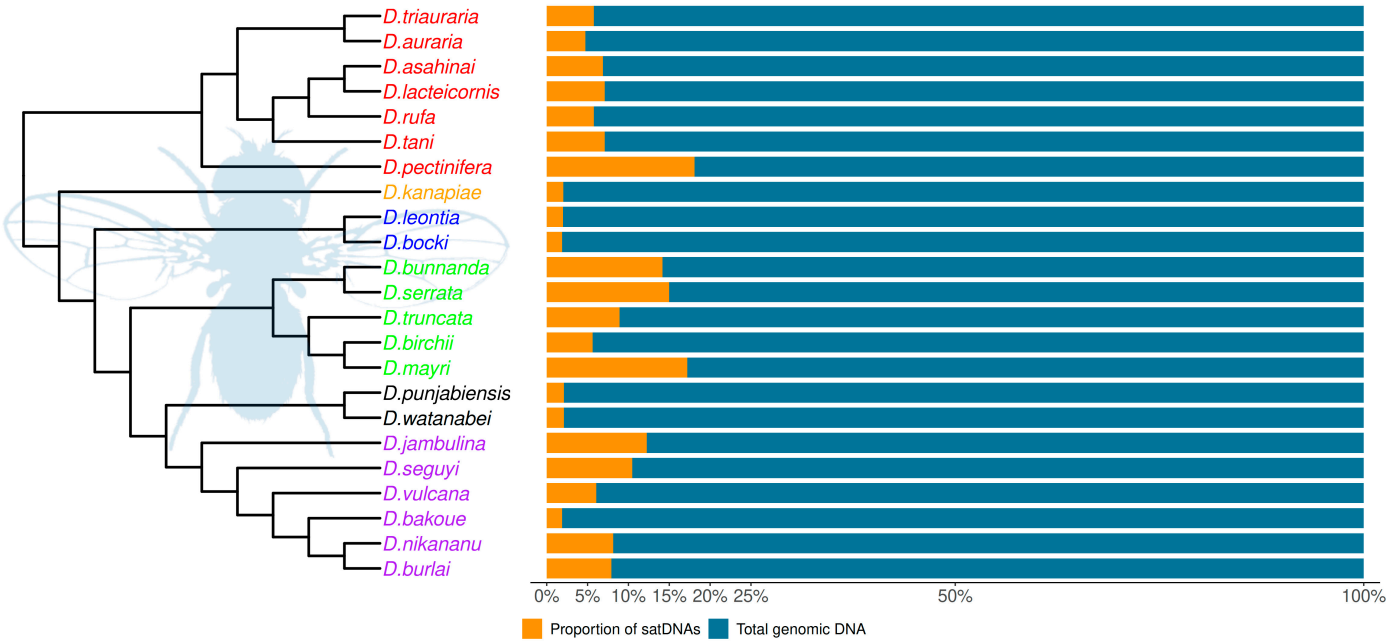

Figure 5. SatDNA genomic proportions in the 23 analyzed species from the *montium* group. The phylogenetic tree was reconstructed according to Conner et al. [30]. Species names are colored according to the subgroups they belong to (see Figure 1).

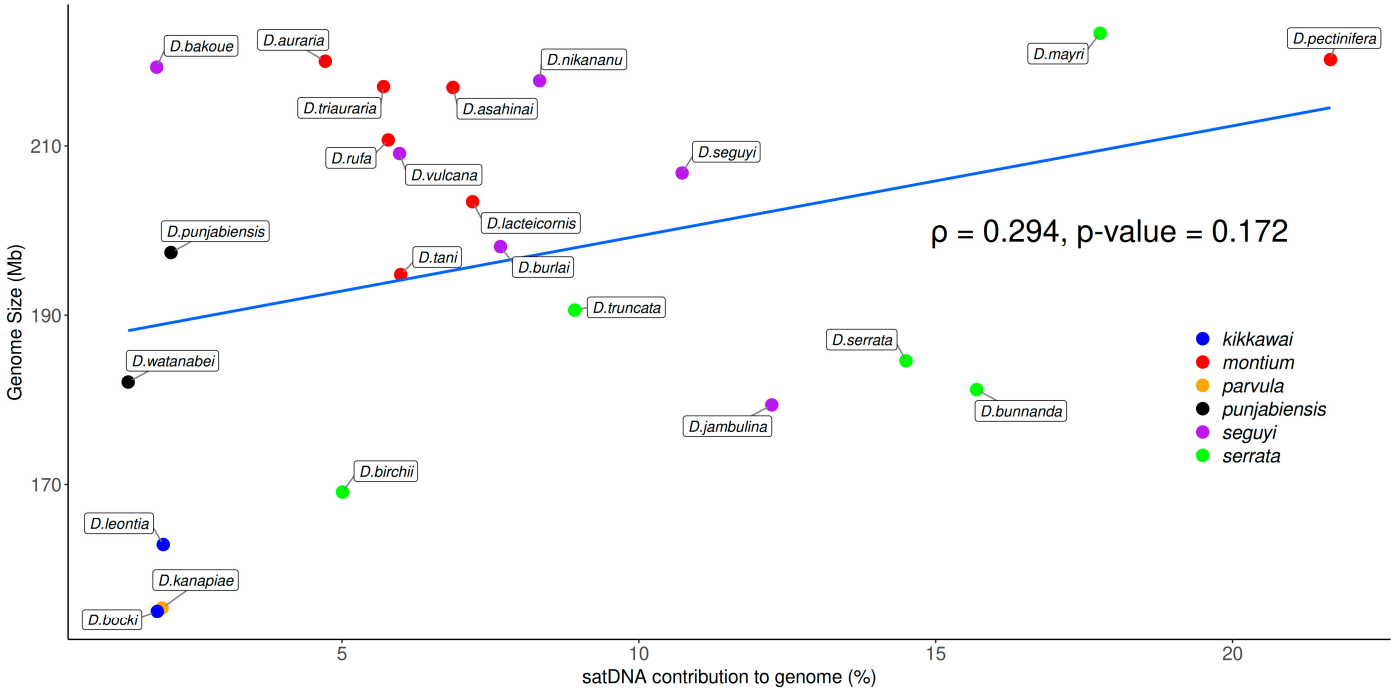

Figure 6. Correlation test between genome size and satDNAs contribution to genome in species from the *montium* group. The *p* value was obtained with Spearman's correlation test.

Considering that Bronski et al. [28] found a strong positive correlation between genome sizes and the whole repetitive DNA content across all the 23 *montium* genomes, we suggest that other repetitive DNAs, probably transposable elements, are the main repetitive DNAs promoting genome size variation in this group of species. In accordance with this hypothesis, a recent study revealed that TE abundance, but not satDNAs, is positively correlated

with genome sizes in *Drosophila* species from the *Sophophora* genus, where species from the *montium* group also belong [49].

3.4. Satellite DNA Distribution across the Montium Phylogeny

Studies in several species of eukaryotes have revealed that satDNAs are among the fastest evolving components of the genome. This assumption is supported by the large number of satDNAs that are found restricted to a few closely related species, or even to a single species [23,26,58].

None of the 101 satDNA families we described here are present in all 23 analyzed species from the *montium* group. This result is not surprising, considering that the common ancestor of the *montium* group lived in Asia more than 19 Mya [29]. In fact, our results showed that most satDNAs families (83%) seem to be restricted to a single species. However, our results obtained with TAREAN do not exclude the possibility that homologous low-copy number, or highly variable repeats, are present in additional species.

From our collection of 101 satDNAs, only 17 are shared by at least two species. The distribution of these 17 satDNAs across the montium group phylogeny is mostly in accordance with the phylogenies proposed by Yassin [29] and Conner et al. [30] at the subgroup level (Figure 7). Several satDNAs are also restricted to species from the same subgroup, such as dmgsat-1, dmgsat-2, and dmgsat-3 from the *montium* subgroup, dmgsatDNA-4 in the *punjabiensis* subgroup, dmgsat-5, dmgsat-6, dmgsat-7, dmgsat-8, dmgsat-9 in the *serrata* subgroup, dmgsat-10, dmgsat-11, and dmgsat-12 in the *kikkawai* subgroup, and dmgsat-14 in species from the *seguyi* subgroup. The remaining satDNAs (dmgsat-13, dmgsat-15, dmgsat-16, and dmgsat-17) are shared between species from different subgroups.

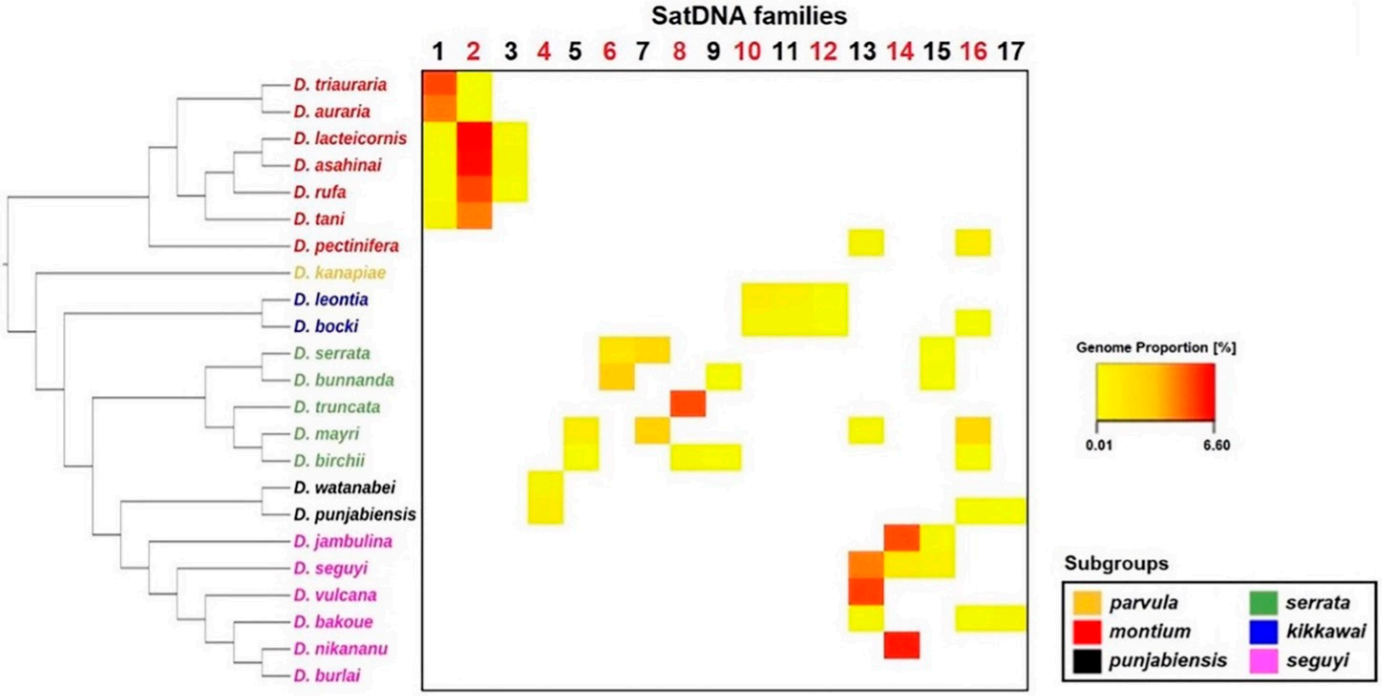

Figure 7. Heatmap showing the genomic proportion for each satDNA family shared by at least two species. The phylogenetic tree was reconstructed according to Conner et al. [30] using the Archaeopteryx software [31]. Species names are colored according to the subgroups they belong to (see Figure 1). The genomic proportion values for each satDNA are described in Table S1.

3.5. Satellite DNA Emergence from DINEs

The *Drosophila* interspersed elements, or DINEs, are abundant (>1000 copies) transposable elements (TEs) found in several *Drosophila* species [59]. They are classified as nonau-

tonomous variants of Helitrons, called Helentrons, and their general structure consist of two conserved blocks (A and B) separated by central tandem repeats (CTRs) (Figure 8A) [60].

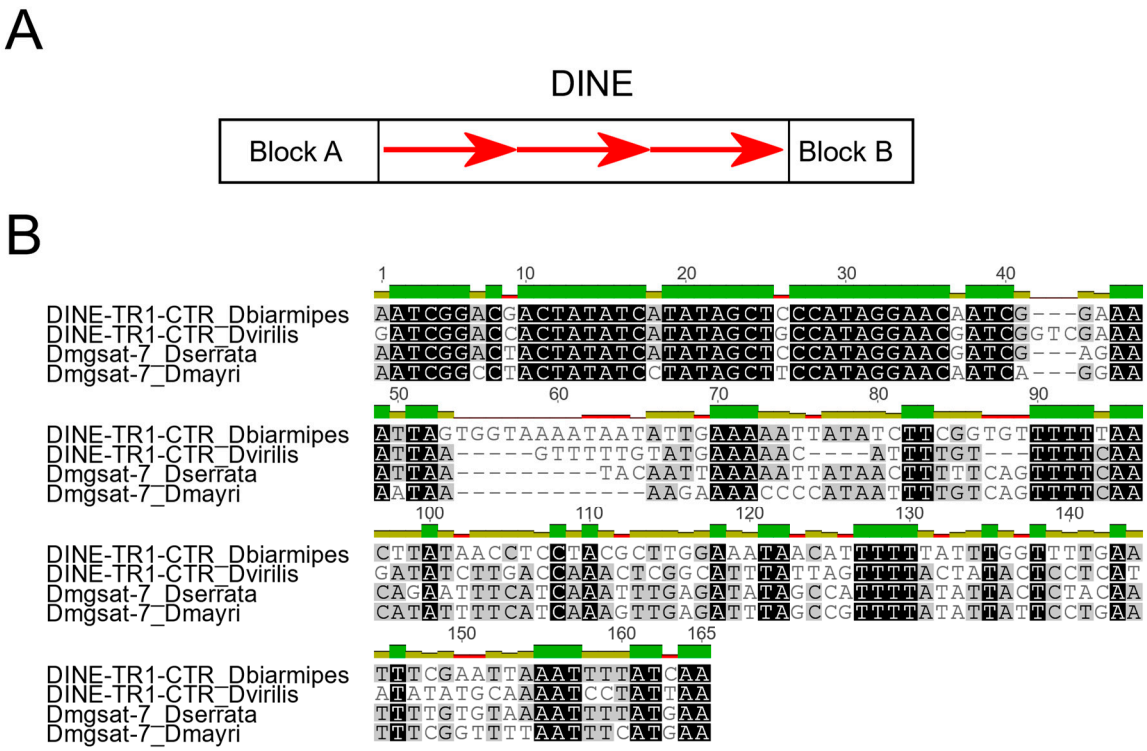

Figure 8. (**A**) General structure of DINEs, including DINE-TR1. Red arrows correspond to the central tandem repeats (CTRs). (**B**) Nucleotide sequence alignment (MUSCLE method) [61] containing DINE-TR1 CTR consensus sequences from D. *biarmipes* and *D. virilis* and the dmgsat-7 consensus sequences from *D. serrata* and *D. mayri*.

Dias et al. [62] identified a DINE variant, named DINE-TR1, to be present in several *Drosophila* species and even outside the genus (in *Bactrocera tryoni*). This DINE-TR1 has CTRs of ~150 bp which are homologous across species. Interestingly, these CTRs have undergone amplification to satDNA-like arrays independently twice across the *Drosophila* phylogeny, both in the ancestral species of *D. virilis* and *D. americana*, and also in *D. biarmipes* [62].

In the present work, we investigated if our collection of 101 satDNA families from the *montium* group shares homology with transposable elements, specially Helitrons. For this purpose, we used our whole collection of satDNA consensus sequences from each satDNA family to screen the CENSOR database on Repbase [63] for homologous known TEs. The results are shown in Table S1. We found that 12 satDNA families (dmgsat-1, dmgsat-7, dmgsat-8, dmgsat-14, dmgsat-20, dmgsat-22, dmgsat-41, dmgsat-67, dmgsat-79, dmgsat-81, dmgsat-84, and dmgsat-91) share regions of DNA sequence identity > 70% to *Drosophila* Helitrons (Table S2). From these satDNAs, we selected dmgsat-7, present in *D. mayri* and *D. serrata* from the *serrata* subgroup, for further in-deep analysis. This was done because its repeat units are very similar in length (~150 bp) to the CTRs present in DINE-TR1. We found that dmgsat-7 consensus sequences are homologous to CTRs present in DINE-TR1 from *D. biarmipes* and *D. virilis,* suggesting that dmgsat-7 is another case of satDNA emergence from DINE-TR1-expanded CTRs (Figure 8B). Interestingly, high sequence identity is limited to the first 30 bp, which possibly indicates the participation of this conserved segment in some functional role, as proposed by Dias et al. [62]. The dmgsat-7 genomic proportion is high in *D. mayri* (2.5%) and *D. serrata* (2.0%) (Figure 7 and Table S1). These values are close to the genomic proportion of the expanded DINE-TR1 CTRs found in *D. virilis* (1.6%) and *D. americana* (2.2%) [43]. To date, *Drosophila serrata* is the

only species from the *serrata* subgroup whose genome has been sequenced with long-read sequencing technology (GenBank: GCA_002093755.1) [64], which allowed us to investigate the size of dmgsat-7 arrays in more detail. Accordingly, we were able to detect dmgsat-7 uninterrupted arrays up to ~ 82.6 kb (~ 540 tandem copies) in *D. serrata* (Table S3).

In summary, our results show that dmgsat-7 is another example of a satDNA derived from the expansion of internal tandem repeats present in DINE-TR1, reinforcing the importance of DINE-TR1 as a potential source for the emergence of satDNAs, as previously suggested [62].

4. Conclusions

With the advent of a new generation DNA sequencing techniques, new bioinformatics tools have been providing efficient ways to identify and classify repetitive DNAs [65,66]. In this context, the TAREAN pipeline was designed as a tool for the identification of satDNA sequences from unassembled short reads. Several studies show that TAREAN is an efficient method for the identification of satDNAs from eukaryotic genomes (e.g., [27,43,45,67,68]).

TAREAN analyses, combined with subsequent manual curation, revealed the presence of 101 satDNAs in 23 *Drosophila* species from the *montium* group, most of them being reported here for the first time. The data presented are expected to provide the framework for future genomic/satellite DNA studies in this group. In particular, the only reported changes in the karyotype configuration of species from the *montium* group concern changes in the amount of heterochromatin [32,33]. It will be interesting to investigate whether these changes are associated with the satDNAs described here.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes14020300/s1, Figure S1: Graph layouts of dmgsat-1 clusters retrieved by TAREAN; Figure S2: Correlation test between genome size and contribution to genome of the 397 initial clusters retrieved by TAREAN (see Figure 2) from the 23 analyzed species from the *montium* group; Table S1: General features of the 101 satDNA families identified in our study; Table S2: SatDNA families in the *montium* group sharing homology with Helitron transposable elements; Table S3: Top ten contigs (sorted by total score) containing copies of dmgsat-7 in *D. serrata*; File S1: Satellite DNA consensus sequences of 101 satellite DNAs identified in species from the *montium* group.

Author Contributions: Conceptualization: B.S.M.L.S. and G.C.S.K.; Methodology: B.S.M.L.S. and G.C.S.K.; Validation: B.S.M.L.S. and G.C.S.K.; Formal Analysis: B.S.M.L.S. and A.C.R.P.; Investigation: B.S.M.L.S.; Resources: G.C.S.K.; Data Curation: B.S.M.L.S. and A.C.R.P.; Writing—Original Draft: B.S.M.L.S.; Preparation: B.S.M.L.S., A.C.R.P. and G.C.S.K.; Writing—Review & Editing: B.S.M.L.S. and G.C.S.K.; Visualization: B.S.M.L.S., A.C.R.P. and G.C.S.K.; Supervision: G.C.S.K.; Project Administration: G.C.S.K.; Funding Acquisition: G.C.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (fellowship 308926/2021-8 to GCSK) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) (fellowship to BSMLS).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Pedro Heringer, Rafaella Soares, Matheus de Morais, and Ana Mattioli for all valuable comments and suggestions to improve our work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. de Koning, A.P.J.; Gu, W.; Castoe, T.A.; Batzer, M.A.; Pollock, D.D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* **2011**, *7*, e1002384. [CrossRef] [PubMed]
2. López-Flores, I.; Garrido-Ramos, M.A. The Repetitive DNA Content of Eukaryotic Genomes. In *Genome Dynamics*; Garrido-Ramos, M.A., Ed.; S. Karger AG: Basel, Switzerland, 2012; Volume 7, pp. 1–28. ISBN 978-3-318-02149-3.
3. Tautz, D. Notes on the Definition and Nomenclature of Tandemly Repetitive DNA Sequences. In *DNA Fingerprinting: State of the Science*; Pena, S.D.J., Chakraborty, R., Epplen, J.T., Jeffreys, A.J., Eds.; Birkhäuser Basel: Basel, Switzerland, 1993; pp. 21–28; ISBN 978-3-7643-2906-8.
4. Charlesworth, B.; Sniegowski, P.; Stephan, W. The Evolutionary Dynamics of Repetitive DNA in Eukaryotes. *Nature* **1994**, *371*, 215–220. [CrossRef] [PubMed]
5. Plohl, M.; Meštrović, N.; Mravinac, B. Satellite DNA Evolution. In *Genome Dynamics*; Garrido-Ramos, M.A., Ed.; S. Karger AG: Basel, Switzerland, 2012; Volume 7, pp. 126–152. ISBN 978-3-318-02149-3.
6. Garrido-Ramos, M. Satellite DNA: An Evolving Topic. *Genes* **2017**, *8*, 230. [CrossRef] [PubMed]
7. Kuhn, G.C.S.; Küttler, H.; Moreira-Filho, O.; Heslop-Harrison, J.S.; Heslop-Harrison, J.S. The 1.688 Repetitive DNA of *Drosophila*: Concerted Evolution at Different Genomic Scales and Association with Genes. *Mol. Biol. Evol.* **2011**, *29*, 7–11. [CrossRef]
8. Brajković, J.; Pezer, Ž.; Bruvo-Mađarić, B.; Sermek, A.; Feliciello, I.; Ugarković, Đ. Dispersion Profiles and Gene Associations of Repetitive DNAs in the Euchromatin of the Beetle *Tribolium castaneum*. *G3 Genes Genomes Genet.* **2018**, *8*, 875–886. [CrossRef]
9. Sproul, J.S.; Khost, D.E.; Eickbush, D.G.; Negm, S.; Wei, X.; Wong, I.; Larracuente, A.M. Dynamic Evolution of Euchromatic Satellites on the X Chromosome in *Drosophila melanogaster* and the *simulans* Clade. *Mol. Biol. Evol.* **2020**, *37*, 2241–2256. [CrossRef]
10. Feliciello, I.; Pezer, Ž.; Sermek, A.; Bruvo Mađarić, B.; Ljubić, S.; Ugarković, Đ. Satellite DNA-Mediated Gene Expression Regulation: Physiological and Evolutionary Implication. In *Satellite DNAs in Physiology and Evolution*; Ugarković, Đ., Ed.; Progress in Molecular and Subcellular Biology; Springer International Publishing: Cham, Switzerland, 2021; Volume 60, pp. 145–167; ISBN 978-3-030-74888-3.
11. Lauria Sneideman, M.P.; Meller, V.H. Drosophila Satellite Repeats at the Intersection of Chromatin, Gene Regulation and Evolution. In *Satellite DNAs in Physiology and Evolution*; Ugarković, Đ., Ed.; Progress in Molecular and Subcellular Biology; Springer International Publishing: Cham, Switzerland, 2021; Volume 60, pp. 1–26. ISBN 978-3-030-74888-3.
12. Pathak, R.; Mamillapalli, A.; Rangaraj, N.; Kumar, R.; Vasanthi, D.; Mishra, K.; Mishra, R.K. AAGAG Repeat RNA Is an Essential Component of Nuclear Matrix in *Drosophila*. *RNA Biol.* **2013**, *10*, 564–571. [CrossRef]
13. Jagannathan, M.; Cummings, R.; Yamashita, Y.M. A Conserved Function for Pericentromeric Satellite DNA. *eLife* **2018**, *7*, e34122. [CrossRef]
14. Jagannathan, M.; Cummings, R.; Yamashita, Y.M. The Modular Mechanism of Chromocenter Formation in *Drosophila*. *eLife* **2019**, *8*, e43938. [CrossRef]
15. Rošić, S.; Köhler, F.; Erhardt, S. Repetitive Centromeric Satellite RNA Is Essential for Kinetochore Formation and Cell Division. *J. Cell Biol.* **2014**, *207*, 335–349. [CrossRef]
16. Louzada, S.; Lopes, M.; Ferreira, D.; Adega, F.; Escudeiro, A.; Gama-Carvalho, M.; Chaves, R. Decoding the Role of Satellite DNA in Genome Architecture and Plasticity—An Evolutionary and Clinical Affair. *Genes* **2020**, *11*, 72. [CrossRef] [PubMed]
17. Flynn, J.M.; Hu, K.B.; Clark, A.G. Three Recent Sex Chromosome-to-Autosome Fusions in a *Drosophila virilis* Strain with High Satellite Content. *bioRxiv* **2021**, 2021-06.
18. Ferree, P.M.; Barbash, D.A. Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in *Drosophila*. *PLoS Biol.* **2009**, *7*, e1000234. [CrossRef]
19. Ugarkovic, D. Functional Elements Residing within Satellite DNAs. *EMBO Rep.* **2005**, *6*, 1035–1039. [CrossRef]
20. Ahmad, S.F.; Singchat, W.; Jehangir, M.; Suntronpong, A.; Panthum, T.; Malaivijitnond, S.; Srikulnath, K. Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. *Cells* **2020**, *9*, 2714. [CrossRef] [PubMed]
21. Strachan, T.; Webb, D.; Dover, G.A. Transition Stages of Molecular Drive in Multiple-Copy DNA Families in *Drosophila*. *EMBO J.* **1985**, *4*, 1701–1708. [CrossRef] [PubMed]
22. Lohe, A.R.; Hilliker, A.J.; Roberts, P.A. Mapping Simple Repeated DNA Sequences in Heterochromatin of *Drosophila melanogaster*. *Genetics* **1993**, *134*, 1149–1174. [CrossRef] [PubMed]
23. Bachmann, L.; Sperlich, D. Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. *Mol. Biol. Evol.* **1993**, *10*, 647–659. [CrossRef]
24. Kuhn, G.C.S. Satellite DNA Transcripts Have Diverse Biological Roles in *Drosophila*. *Heredity* **2015**, *115*, 1–2. [CrossRef]
25. Khost, D.E.; Eickbush, D.G.; Larracuente, A.M. Single-Molecule Sequencing Resolves the Detailed Structure of Complex Satellite DNA Loci in *Drosophila melanogaster*. *Genome Res.* **2017**, *27*, 709–721. [CrossRef]
26. Kuhn, G.C.S.; Heringer, P.; Dias, G.B. Structure, Organization, and Evolution of Satellite DNAs: Insights from the *Drosophila repleta* and *D. virilis* Species Groups. In *Satellite DNAs in Physiology and Evolution*; Ugarković, Đ., Ed.; Progress in Molecular and Subcellular Biology; Springer International Publishing: Cham, Switzerland, 2021; Volume 60, pp. 27–56. ISBN 978-3-030-74888-3.
27. Novák, P.; Robledillo, L.Á.; Koblížková, A.; Vrbová, I.; Neumann, P.; Macas, J. TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **2017**, *45*, e111. [CrossRef] [PubMed]

28. Bronski, M.J.; Martinez, C.C.; Weld, H.A.; Eisen, M.B. Whole Genome Sequences of 23 Species from the *Drosophila montium* Species Group (Diptera: Drosophilidae): A Resource for Testing Evolutionary Hypotheses. *G3 Genes Genomes Genet.* **2020**, *10*, 1443–1455. [CrossRef] [PubMed]
29. Yassin, A. Phylogenetic Biogeography and Classification of the *Drosophila montium* Species Group (Diptera: Drosophilidae). *Ann. Société Entomol. Fr. (N.S.)* **2018**, *54*, 167–175. [CrossRef]
30. Conner, W.R.; Delaney, E.K.; Bronski, M.J.; Ginsberg, P.S.; Wheeler, T.B.; Richardson, K.M.; Peckenpaugh, B.; Kim, K.J.; Watada, M.; Hoffmann, A.A.; et al. A Phylogeny for the *Drosophila montium* Species Group: A Model Clade for Comparative Analyses. *Mol. Phylogenetics Evol.* **2021**, *158*, 107061. [CrossRef] [PubMed]
31. Han, M.V.; Zmasek, C.M. PhyloXML: XML for Evolutionary Biology and Comparative Genomics. *BMC Bioinform.* **2009**, *10*, 356. [CrossRef]
32. Baimai, V. Metaphase Karyotypes of Certain Species of the *Drosophila montium* Subgroup. *JPN J. Genet.* **1980**, *55*, 165–175. [CrossRef]
33. Venkat, S.; Ranganath, R.A. Localization and Characterization of Heterochromatin among Four Species of the *montium* Subgroup of *Drosophila*. *Cytologia* **2007**, *72*, 279–286. [CrossRef]
34. Ma, J.; Jackson, S.A. Retrotransposon Accumulation and Satellite Amplification Mediated by Segmental Duplication Facilitate Centromere Expansion in Rice. *Genome Res.* **2006**, *16*, 251–259. [CrossRef] [PubMed]
35. Kuhn, G.C.S.; Teo, C.H.; Schwarzacher, T.; Heslop-Harrison, J.S. Evolutionary Dynamics and Sites of Illegitimate Recombination Revealed in the Interspersion and Sequence Junctions of Two Nonhomologous Satellite DNAs in Cactophilic *Drosophila* Species. *Heredity* **2009**, *102*, 453–464. [CrossRef]
36. Wei, K.H.-C.; Grenier, J.K.; Barbash, D.A.; Clark, A.G. Correlated Variation and Population Differentiation in Satellite DNA Abundance among Lines of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 18793–18798. [CrossRef]
37. Ávila Robledillo, L.; Koblížková, A.; Novák, P.; Böttinger, K.; Vrbová, I.; Neumann, P.; Schubert, I.; Macas, J. Satellite DNA in *Vicia faba* Is Characterized by Remarkable Diversity in Its Sequence Composition, Association with Centromeres, and Replication Timing. *Sci. Rep.* **2018**, *8*, 5838. [CrossRef] [PubMed]
38. Palacios-Gimenez, O.M.; Koelman, J.; Palmada-Flores, M.; Bradford, T.M.; Jones, K.K.; Cooper, S.J.B.; Kawakami, T.; Suh, A. Comparative Analysis of Morabine Grasshopper Genomes Reveals Highly Abundant Transposable Elements and Rapidly Proliferating Satellite DNA Repeats. *BMC Biol.* **2020**, *18*, 199. [CrossRef] [PubMed]
39. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef] [PubMed]
40. Novák, P.; Neumann, P.; Macas, J. Global Analysis of Repetitive DNA from Unassembled Sequence Reads Using RepeatExplorer2. *Nat. Protoc.* **2020**, *15*, 3745–3776. [CrossRef]
41. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. *Bioinformatics* **2012**, *28*, 1647–1649. [CrossRef] [PubMed]
42. de Lima, L.G.; Svartman, M.; Kuhn, G.C.S. Dissecting the Satellite DNA Landscape in Three Cactophilic *Drosophila* Sequenced Genomes. *G3 Genes Genomes Genet.* **2017**, *7*, 2831–2843. [CrossRef]
43. Silva, B.S.M.L.; Heringer, P.; Dias, G.B.; Svartman, M.; Kuhn, G.C.S. De Novo Identification of Satellite DNAs in the Sequenced Genomes of *Drosophila virilis* and *D. americana* Using the RepeatExplorer and TAREAN Pipelines. *PLoS ONE* **2019**, *14*, e0223466. [CrossRef]
44. da Silva, M.J.; Fogarin Destro, R.; Gazoni, T.; Narimatsu, H.; Pereira dos Santos, P.S.; Haddad, C.F.B.; Parise-Maltempi, P.P. Great Abundance of Satellite DNA in *Proceratophrys* (Anura, Odontophrynidae) Revealed by Genome Sequencing. *Cytogenet. Genome Res.* **2020**, *160*, 141–147. [CrossRef]
45. Sena, R.S.; Heringer, P.; Valeri, M.P.; Pereira, V.S.; Kuhn, G.C.S.; Svartman, M. Identification and Characterization of Satellite DNAs in Two-Toed Sloths of the Genus *Choloepus* (Megalonychidae, Xenarthra). *Sci. Rep.* **2020**, *10*, 19202. [CrossRef]
46. DiBartolomeis, S.M.; Tartof, K.D.; Jackson, F.R. A superfamily of *Drosophila* satellite related (SR) DNA repeats restricted to the X chromosome euchromatin. *Nucleic Acids Res.* **1992**, *20*, 1113–1116. [CrossRef]
47. Lohe, A.R.; Brutlag, D.L. Adjacent Satellite DNA Segments in *Drosophila*. *J. Mol. Biol.* **1987**, *194*, 171–179. [CrossRef] [PubMed]
48. Jagannathan, M.; Warsinger-Pepe, N.; Watase, G.J.; Yamashita, Y.M. Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3 Genes Genomes Genet.* **2017**, *7*, 693–704. [CrossRef]
49. de Lima, L.G.; Ruiz-Ruano, F.J. In-Depth Satellitome Analyses of 37 *Drosophila* Species Illuminate Repetitive DNA Evolution in the *Drosophila* Genus. *Genome Biol. Evol.* **2022**, *14*, evac064. [CrossRef]
50. Palomeque, T.; Lorite, P. Satellite DNA in Insects: A Review. *Heredity* **2008**, *100*, 564–573. [CrossRef] [PubMed]
51. Melters, D.P.; Bradnam, K.R.; Young, H.A.; Telis, N.; May, M.R.; Ruby, J.; Sebra, R.; Peluso, P.; Eid, J.; Rank, D.; et al. Comparative Analysis of Tandem Repeats from Hundreds of Species Reveals Unique Insights into Centromere Evolution. *Genome Biol.* **2013**, *14*, R10. [CrossRef]
52. Gall, J.; Cohen, E.; Polan, M. Repetitive DNA Sequences in *Drosophila*. *Chromosoma* **1971**, *33*, 319–344. [CrossRef] [PubMed]
53. Gall, J.G.; Atherton, D.D. Satellite DNA Sequences in *Drosophila virilis*. *J. Mol. Biol.* **1974**, *85*, 633–664. [CrossRef] [PubMed]

54. Schmidt, E.R. Two AT-Rich Satellite DNAs in the Chironomid *Glyptotendipes barbipes* (Staeger): Isolation and Localization in Polytene Chromosomes of *G. barbipes* and *Chironomus thummi*. *Chromosoma* **1980**, *79*, 315–328. [CrossRef]
55. Ganal, M.; Hemleben, V. Different AT-Rich Satellite DNAs in *Cucurbita pepo* and *Cucurbita maxima*. *Theoret. Appl. Genetics* **1986**, *73*, 129–135. [CrossRef]
56. Bosco, G.; Campbell, P.; Leiva-Neto, J.T.; Markow, T.A. Analysis of *Drosophila* Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species. *Genetics* **2007**, *177*, 1277–1290. [CrossRef] [PubMed]
57. Craddock, E.M.; Gall, J.G.; Jonas, M. Hawaiian *Drosophila* Genomes: Size Variation and Evolutionary Expansions. *Genetica* **2016**, *144*, 107–124. [CrossRef]
58. Bachmann, L.; Venanzetti, F.; Sbordoni, V. Characterization of a Species-Specific Satellite DNA Family of *Dolichopoda schiavazzii* (Orthoptera, Rhaphidophoridae) Cave Crickets. *J. Mol. Evol.* **1994**, *39*, 274–281. [CrossRef] [PubMed]
59. Yang, H.-P.; Barbash, D.A. Abundant and Species-Specific DINE-1 Transposable Elements in 12 *Drosophila* Genomes. *Genome Biol.* **2008**, *9*, R39. [CrossRef] [PubMed]
60. Thomas, J.; Pritham, E.J. *Helitrons*, the Eukaryotic Rolling-Circle Transposable Elements. *Microbiol. Spectr.* **2015**, *3*, 3–4. [CrossRef] [PubMed]
61. Edgar, R.C. MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinform.* **2004**, *5*, 113. [CrossRef]
62. Dias, G.B.; Heringer, P.; Svartman, M.; Kuhn, G.C.S. Helitrons Shaping the Genomic Architecture of *Drosophila*: Enrichment of DINE-TR1 in α- and β-Heterochromatin, Satellite DNA Emergence, and PiRNA Expression. *Chromosome. Res.* **2015**, *23*, 597–613. [CrossRef] [PubMed]
63. Kohany, O.; Gentles, A.J.; Hankus, L.; Jurka, J. Annotation, Submission and Screening of Repetitive Elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinform.* **2006**, *7*, 474. [CrossRef]
64. Allen, S.L.; Delaney, E.K.; Kopp, A.; Chenoweth, S.F. Single-Molecule Sequencing of the *Drosophila serrata* Genome. *G3 Genes Genomes Genet.* **2017**, *7*, 781–788. [CrossRef]
65. Dias, G.B.; Svartman, M.; Delprat, A.; Ruiz, A.; Kuhn, G.C.S. Tetris Is a Foldback Transposon That Provided the Building Blocks for an Emerging Satellite DNA of *Drosophila virilis*. *Genome Biol. Evol.* **2014**, *6*, 1302–1313. [CrossRef]
66. Dias, C.A.R.; Kuhn, G.C.S.; Svartman, M.; dos Santos Júnior, J.E.; Santos, F.R.; Pinto, C.M.; Perini, F.A. Identification and Characterization of Repetitive DNA in the Genus *Didelphis* Linnaeus, 1758 (Didelphimorphia, Didelphidae) and the Use of Satellite DNAs as Phylogenetic Markers. *Genet. Mol. Biol.* **2021**, *44*, e20200384. [CrossRef] [PubMed]
67. Valeri, M.P.; Dias, G.B.; do Espírito Santo, A.A.; Moreira, C.N.; Yonenaga-Yassuda, Y.; Sommer, I.B.; Kuhn, G.C.S.; Svartman, M. First Description of a Satellite DNA in Manatees' Centromeric Regions. *Front. Genet.* **2021**, *12*, 694866. [CrossRef] [PubMed]
68. Gasparotto, A.E.; Milani, D.; Martí, E.; Ferretti, A.B.S.M.; Bardella, V.B.; Hickmann, F.; Zrzavá, M.; Marec, F.; Cabral-de-Mello, D.C. A Step Forward in the Genome Characterization of the Sugarcane Borer, *Diatraea saccharalis*: Karyotype Analysis, Sex Chromosome System and Repetitive DNAs through a Cytogenomic Approach. *Chromosoma* **2022**, *131*, 253–267. [CrossRef] [PubMed]