Dynamic evolution of euchromatic satellites on the X chromosome in *Drosophila melanogaster* and the *simulans* clade

Sproul, J.S.*[1], Khost, D.E.* [1], Eickbush, D.G. [1], Negm, S. [1], Wei, X. [2], Wong, I. [1], and A.M. Larracuente[1]

*these authors have equal contribution

1. University of Rochester, Department of Biology, 337 Hutchison Hall, Rochester, NY, 14627
2. Department of Biomedical Genetics, University of Rochester Medical Center, 601 Elmwood Ave. Rochester, NY, 14642

Corresponding author: alarracu@bio.rochester.edu

Orcids: JSS 0000-0002-6747-3537, WX 0000-0001-9952-3757, AML 0000-0001-5944-5686

Keywords: satellite DNA, Drosophila, genome evolution, repeats, eccDNA

Running title: Dynamic evolution of euchromatic satellite repeats.

ABSTRACT

Repeats are abundant in eukaryotic genomes and contribute to differences in genome size and organization among organisms. The large blocks of tandem repeats—satellite DNAs (satDNAs)—frequently found in regions of low recombination can turn over rapidly between species, with potential consequences for genome evolution and speciation. Short blocks of satDNA also exist in the euchromatin, where they are particularly abundant on the X chromosome. These euchromatic repeats can affect gene expression and some have roles dosage compensation. Despite their abundance and impact on important phenotypes, we know little about the detailed evolutionary dynamics and the processes that shape satDNA distributions in genomes over short evolutionary time scales. Here we use high-quality genome assemblies to study the evolutionary dynamics of satDNA across closely related species: *Drosophila melanogaster* and three species of the *simulans* clade (*D. simulans*, *D. sechellia*, and *D. mauritiana*). We focus on two complex satDNA families, *Rsp-like* and *1.688* gm/cm3. These repeats are highly dynamic in the heterochromatin, where their genomic location varies. We discovered that euchromatic repeats are similarly dynamic, changing in abundance, number of clusters, and composition within clusters, even across the *simulans* clade. While *1.688* is an old repeat family, *Rsp-like* has recently proliferated, spreading to new genomic locations across the X chromosome independently in *D. simulans* and *D. mauritiana*. We infer that extrachromosomal circular DNA integration and/or interlocus gene conversions resolved by microhomology-mediated repair pathways could account for satDNA proliferation in genomes. The divergence of repeat landscapes between species may have important consequences for genome evolution.

BACKGROUND

Eukaryotic genomes are replete with large blocks of tandemly repeated DNA sequences. Named for their distinct "satellite" bands on cesium chloride gradients [1-3], these so-called satellite DNAs (satDNA) can comprise large fractions of eukaryotic genomes [4, 5]. SatDNAs are a major component of heterochromatin—large blocks of satDNAs accumulate near centromeres and telomeres in many organisms [6, 7]. The location, abundance, and sequence of these heterochromatic satDNAs can turnover rapidly [5, 8] creating divergent repeat profiles between species [9]. The rapid evolution of satDNA can have broad evolutionary consequences due to its role in diverse processes including chromatin packaging [10] and chromosome segregation [11]. For example, variation in satDNA can impact centromere location and stability [12], meiotic drive systems [13-15], hybrid incompatibilities [16], and genome evolution [4, 17, 18].

Novel satDNAs can arise from the amplification of unique sequences through replication slippage [19, 20], unequal exchange, rolling circle replication [4, 21-23], or even from transposable elements (TEs) [24-26]. However, much of the species-level differences in satDNA arises through movement and divergence of ancestral satellites inherited through common decent [27]. Unequal exchange between different repeats within a tandem array can lead to expansions and contractions of repeats at a locus [28], and along with gene conversion, lead to the homogenization of repeated sequences within species—both within repeat arrays (*e.g.,* [29]) and between repeats on different chromosomes— and the divergence of repeats between species (reviewed in [30]). These processes result in the concerted evolution [31] of satDNAs [9] and of multicopy gene families like rDNA and histones [32], leading to species-specific repeat profiles.

While large blocks of satDNAs accumulate in heterochromatin, small blocks of tandem repeats occur in euchromatic regions of the genome and are particularly enriched on X chromosomes [33, 34]. Some euchromatic X-linked repeats have sequence similarity to the large blocks of heterochromatic satDNAs (*e.g.,* [33-35]). Recent studies suggest that these repeats may play roles in gene regulation, chromatin regulation, and X chromosome recognition. Some satDNA repeats occur in or near genes, where they may act as "evolutionary tuning knobs" on gene expression [36]. Consistent with this hypothesis, a dispersed satDNA in red flour beetles [37] can alter nearby gene expression following heat shock by modifying local chromatin state [38]. Beyond their effects on local gene expression, euchromatic satDNAs may also aid in chromosome recognition. To compensate for differences in X-linked gene dosage between males and females, the Male Specific Lethal (MSL) complex binds to the X chromosome and upregulates gene expression ~2-fold in *Drosophila melanogaster* males [39]. Small blocks of the *1.688* satellite across the X euchromatin [33, 35] may contribute to X chromosome recognition by MSL through a siRNA-mediated mechanism [40-42]. Similarly, euchromatic X-linked *1.688* satellite clusters interact with another chromosome-targeting protein called Painting of Fourth (POF) [43, 44].

The precise mechanisms underlying the rapid expansion, movement, and rearrangement of satDNAs across the genome are not well understood. Recombination-based mechanisms can cause local rearrangements or large-scale structural rearrangements such as chromosomal translocations [45, 46]. Intra-chromatid recombination events give rise to extrachromosomal circular DNAs (eccDNAs) that are common across eukaryotic organisms [47-53] and may be produced in abundance under conditions of stress or during aging. These eccDNAs may contribute to the rapidly changing repeat landscape across genomes. Euchromatic repeats also undergo concerted evolution (*e.g.*, [35]) and can evolve rapidly [54], with unknown consequences on genome function. Mechanisms underlying the evolution of euchromatic satDNAs are understudied, in part due to the fact that repeats present challenges to sequence-based and molecular biology approaches.

Here we compare the repeat landscape of *D. melanogaster* and species of the *simulans* clade—*D. simulans, D. mauritiana,* and *D. sechellia*—using new reference genomes based on long single-molecule sequence reads [55]. We focus on two abundant satellite repeat families: *1.688 gm/cm3* and *Rsp-like*. *1.688 g/cm3* (hereafter called *1.688*) is a family of several related repeats named after their monomer lengths, including *260bp, 353bp, 356bp, 359bp,* and *360bp* [56, 57]. *Rsp-like* is a 160-bp repeat named for its similarity to the 120-bp *Responder (Rsp)* satellite [58]. We show that these repeats are highly dynamic in their chromosomal location in both heterochromatin and euchromatin, with compositional shifts in repeat types occurring even between sister species of the *simulans* clade. The *Rsp-like* repeats recently proliferated across the *D. simulans* and *D. mauritiana* X euchromatin. Our results suggest that microhomology-mediated repair events can create novel associations between unrelated satDNA repeat types, which can facilitate their rapid spread across large physical distances in the genome.

RESULTS

SatDNA composition varies across species

Large blocks of satDNA in the heterochromatin change locations on short evolutionary timescales in *Drosophila* species (*e.g.,* [22, 58-60]). To determine the genomic locations of *1.688* and *Rsp-like* satellites in the fly strains for which we have high quality PacBio assemblies [55], we examined their distribution on mitotic chromosomes with fluorescence *in situ* hybridization (FISH). The genomic location of these satDNAs in the heterochromatin varies among species (Fig. 1). Large heterochromatic blocks of *1.688* repeats are primarily X-linked in *D. melanogaster* (*359bp*) and *D. sechellia* but autosomal in *D. simulans* and *D. mauritiana* (Fig. 1). *D. melanogaster* has two smaller blocks of *1.688* family repeats in the heterochromatin of chromosome 3 (*e.g., 356bp* and *353bp*)[56]. The distribution of the *Rsp-like* family is similarly dynamic in the heterochromatin: large blocks are X-linked in *D. simulans,* autosomal in *D. sechellia* (chromosome 2 and 3), and lacking in the heterochromatin of *D. mauritiana* and *D. melanogaster* ([58]; Fig. 1). Although *D. melanogaster* lacks heterochromatic *Rsp-like* repeats, it has a distantly related heterochromatic satellite (*Rsp*) on chromosome 2 [61, 62]. At this broad scale, the rapid turnover of these pericentromeric satDNAs among species is similar to the dynamic turnover of other pericentromeric satellites reported in a wide range of taxa (*e.g.*, [60, 63-66]).

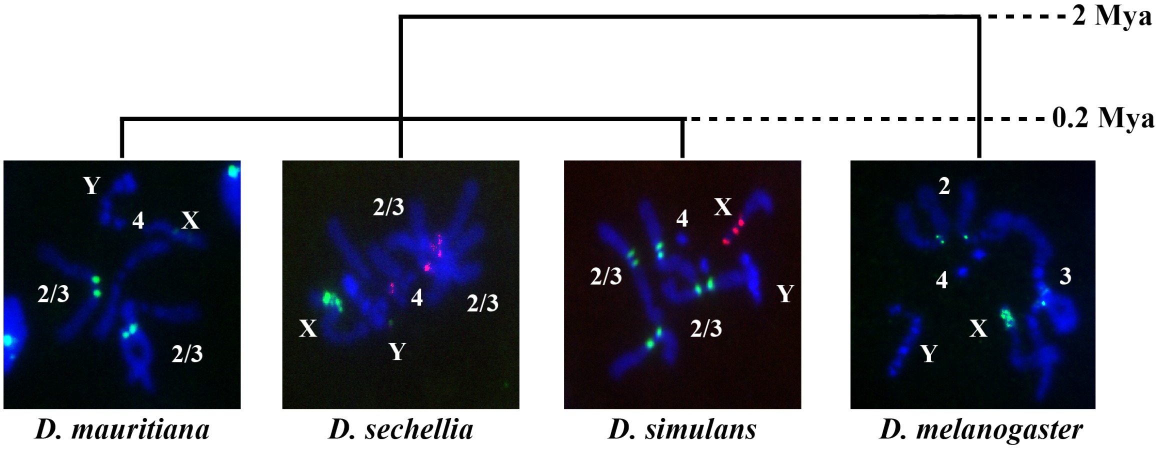


Figure 1. Complex satellites in the heterochromatin of *D. melanogaster* and the *simulans* clade are in different locations. FISH images of mitotic chromosomes showing *Rsp-like* (red) and *1.688* (green) satellites. Chromosomes are counterstained using DAPI.

The *1.688* repeat family also exists in the euchromatin [33-35, 54], where they are over-represented on the X chromosome relative to the autosomes in the *Drosophila* species studied here [55]. We find that *Rsp-like* repeats also exist in the euchromatin. Our annotation of the assemblies shows that these satellites are non-randomly distributed across the X euchromatin. Both satellites accumulate near the telomere (cytoband 1) and in the middle of the X chromosome but are uncommon proximal to cytoband 14 (Figs. 2, S2). We describe the location of these repeats relative to their cytological divisions (*i.e.* cytobands) on *D. melanogaster* polytene chromosomes and hereafter use the terms ‘cytobands’, ‘clusters’, and ‘monomers’ as illustrated in Fig. 2a. We confirmed the euchromatic enrichment of these repeats using FISH on polytene chromosomes, where we see a high density of bands on the polytenized arm of the X chromosome in the *simulans* clade species (*e.g.,* representative FISH image; Fig. S1).

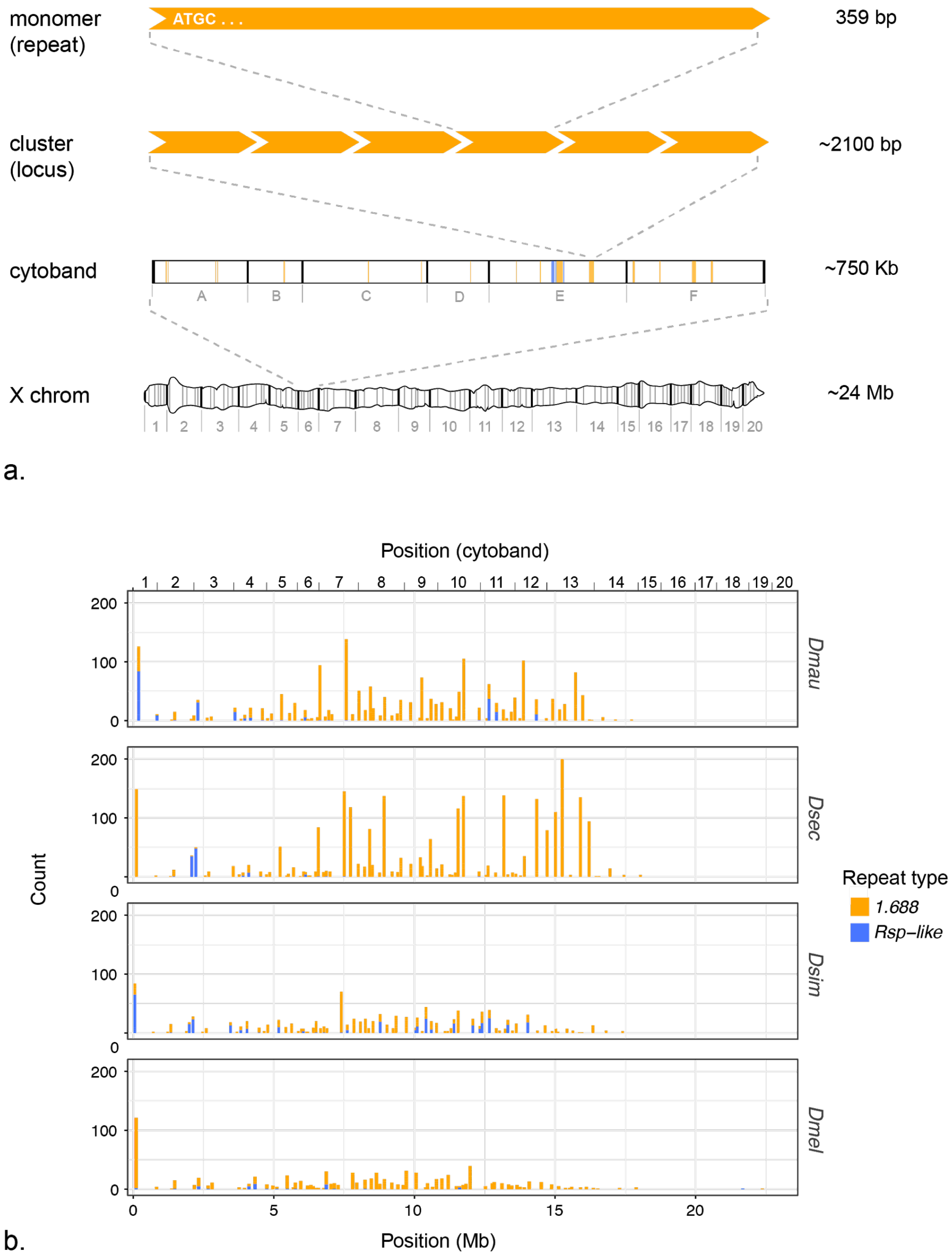


Figure 2: Euchromatic X-linked satellites are unevenly distributed across the X chromosome. (a.) A schematic illustrating terms frequently used in the text. We use ‘cytoband’ to reference large regions of the X chromosome that are defined by banding patterns in polytene chromosomes. We use ‘cluster’ to mean any distinct genomic locus

containing the repeat of interest; typically, clusters contain several tandem repeats, although single-repeat clusters also exist. 'Monomer' refers to a single repeat unit; the example shown represents a *1.688* monomer. (b.**)** The x-axis shows position of *1.688* and *Rsp-like* satDNA clusters along the X chromosome. Counts shown on the y-axis indicate the number of repeat copies (*i.e.*, monomers) within a cluster. Each bar on the chart represents a cytological subdivision (*e.g.*, 1A, 1B, etc.) in which counts of all repeats are pooled.

The abundance of euchromatic complex satellite repeats shows >3-fold variation among species. *D. sechellia* has the most euchromatic X-linked repeats (2588 annotations), followed by *D. mauritiana* (1390) and *D. simulans* (1112), and *D. melanogaster* (849) (Table 1). The high number of repeats in *D. sechellia* may be due to the reduced efficacy of natural selection in this island endemic species, which has a historically low effective population size [67, 68]. The *D. sechellia* X chromosome assembly contains 19 gaps, six of which occur within satellite loci, therefore the X-linked copy number represents a minimum estimate for this species [55].

Table 1. Summary of euchromatic satDNA cluster sizes on X chromosome. Total #: number of total repeats. # clust: total number of clusters at distinct loci. % N=1: percentage of singletons (clusters of a single repeat). % N<4: percentage of small clusters (less than four repeats).

Species	**Total# *1.688***	**# *1.688* clust**	**%N=1 *1.688***	**% N<4 *1.688***	**# *Rsp-like***	**# *Rsp-like* clust**	**% N=1 *Rsp-like***	**% N<4 *Rsp-like***
D. mauritiana	1165	325	24.00	68.31	225	26	30.77	34.62
D. sechellia	2486	308	33.44	82.14	102	12	50.00	58.33
D. simulans	786	324	31.17	89.20	326	38	18.42	34.21
D. melanogaster	808	274	33.94	83.94	41	19	73.68	78.95

Across all species, *1.688* is more abundant than *Rsp-like*, both in terms of total repeats (*i.e.,* the number of euchromatic repeat monomers annotated in our assemblies, Fig. 2a) and the number of clusters (*i.e.*, the number of distinct genomic loci containing repeats, Fig. 2a); however, the majority of *1.688* clusters are small (*i.e.*, contain <4 repeats). This contrasts with *Rsp-like,* where clusters are less abundant but larger on average (Table 1, Fig. S3). Single-monomer clusters exist in both satDNA types; however, they are much more common in *1.688* where they account for ~30% of all clusters (Table 1, Fig. S3). We consider single-monomer clusters to be "dead" as they cannot undergo unequal exchange and expand [7, 30, 69].

The number of total repeats and the number of clusters for each satellite also varies among species. *Rsp-like* shows up to an 8-fold difference in total repeat number and >3-fold difference in number of clusters among species, with *D. simulans* and *D. mauritiana* having more total repeats as well as more clusters compared to *D. sechellia* and *D. melanogaster* (Table 1, Fig. S3). In *D. simulans* and *D. mauritiana, Rsp-like* clusters have apparently spread to cytobands that lack such clusters in one, or both of the other species (*e.g.*, those clusters at cytobands 7-12 in *D. simulans,* and cytobands 11 and 12 in *D. mauritiana*) (Figs. 2, S2). In addition, *D. simulans* and *D. mauritiana* have a lower proportion of single repeat, or 'dead' clusters (18.4% and 30.8%, respectively) than the other species (Table 1). In *1.688, D. sechellia* shows as much as a 3-fold increase in total repeats despite having fewer *1.688* loci than the other *simulans* clade species, a pattern driven by a high number of large clusters (≥50 monomers) in *D. sechellia* (16 clusters ≥50), which are less common in other species (six clusters in *D. mauritiana*, one in *D. simulans* and *D. melanogaster*) (Table 1).

These patterns suggest dynamic turnover of satDNA repeats across the X chromosome euchromatin over short evolutionary time scales. While it is tempting to make a sweeping statement based on these numbers, it is difficult to systematically identify orthologous loci across the X chromosome to accurately quantify the turnover on a locus-by-locus basis. However, we can explore the dynamics of specific clusters for which synteny of unique flanking sequences strongly suggests orthology across species. One such representative cluster is embedded between two genes—*echinus* and *roX1*—at cytoband 3F (Fig. 3). In *D. melanogaster*, this cluster has only two *1.688* repeats, the first of which is truncated, plus an unannotated adjacent region that contains degenerated *1.688* sequence. *D. sechellia* also has *1.688* at this location, but the cluster is expanded relative to *D. melanogaster*. In contrast, both *Rsp-like* and *1.688* repeats are present at this locus in *D. mauritiana* and *D. simulans*; however, each species shows differences in repeat number of the respective satellites (Fig. 3). The *Rsp-like* repeats in *D. mauritiana* and *D. simulans* are homogenized within the locus and are highly divergent between species. The major differences in euchromatic satellite composition among species suggest that euchromatic satellites, like heterochromatic satellites, evolve dynamically over short evolutionary time scales.

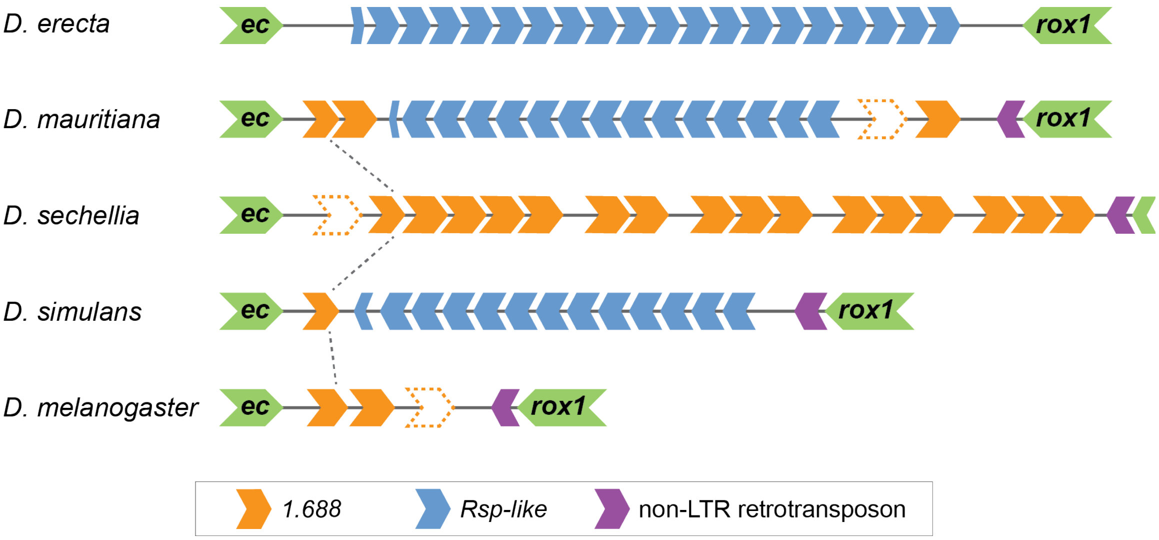


Figure 3: Organization of cytoband 3F repeat cluster. Schematic of 3F cluster in *D. melanogaster* and the *simulans* clade, as well as the outgroup species *D. erecta.* Cluster is flanked by two genes, *echinus* and *roX1* (light green chevrons), with a TE insertion at the distal side of the locus (purple chevrons). Complex satellite monomers are indicated by blue (*Rsp-like*) or orange (*1.688*) chevrons. Chevrons with dotted outline indicate sequences that were not annotated, but were determined manually by BLAST to be highly degenerated satellite monomers. Black dotted lines between species indicate shared repeats.

Association between 1.688 *and* Rsp-like *repeats*

Analysis of the nearest upstream and downstream genomic features relative to *1.688* and *Rsp-like* satellites showed that *Rsp-like* clusters have a non-random distribution, particularly in *D. simulans* and *D. mauritiana*. *Rsp-like* clusters are directly adjacent to, or interspersed with, *1.688* clusters in 82% of euchromatic X-linked clusters in *D. simulans* and in 62% of clusters in *D. mauritiana* (Table 2, Figs. S4–S5). Conversely, the

1.688 clusters do not seem to preferentially associate with *Rsp-like*, though they are often located near genes [35] (Figs. S4–S5).

Table 2: *Rsp-like* clusters associate with *1.688.* # *Rsp-like*: number of *Rsp-like* clusters on X chromosome. #*Rsp-like* / *1.688*: number of *Rsp-like* clusters (including singletons) that have *1.688* repeats within 100bp either upstream or downstream.

Species	***Rsp-like***	***Rsp-like / 1.688***	**% *Rsp-like / 1.688***
D. mauritiana	26	16	62
D. sechellia	12	3	25
D. simulans	38	31	82
D. melanogaster	19	7	37

Evolutionary relationship of satDNAs within and among species

Examination of within-species and all-species phylogenetic trees of satellite repeats led to four major findings. (1) Heterochromatic repeats form clades that are generally separate from euchromatic repeats for both satellites in all species with the exception of *D. sechellia*, for which euchromatic and heterochromatic repeats are interspersed in both *1.688* and *Rsp-like* (Figs. S6–13). (2) *D. sechellia* and *D. mauritiana* (especially the former) show repeated evidence of intralocus expansion of repeats (Figs. S14–15). (3) *1.688* euchromatic repeats have a relatively old diversification history that largely predates the speciation events that gave rise to the study species (Figs. 4–5, S6, S8, S10, S12, S14–15). This contrasts with *Rsp-like* euchromatic repeats, which show evidence of relatively recent diversification, particularly in the *simulans* clade species (Figs. 4–5, S7, S9, S11, S13, S16–17). (4) *Rsp-like* repeats show evidence of two major expansions that occurred in the recent history of the *simulans* clade (Figs. 4–5, S7, S9, S11, S16–17), where the new repeats span large physical distances across the X chromosome (*i.e.*, 'interlocus' expansions) and occurred largely independently in *D. simulans* and *D. mauritiana*. The latter two findings warrant further explanation of evidence lending to their support.

With regard to finding three, within-species trees show contrasting patterns of branch length, nodal support, and local differentiation of repeats, which all indicate a more recent history of *Rsp-like* diversification (Figs. 4–5, S6–12; Supplemental Results; Table S1). The *1.688* all-species tree supports the conclusion that *1.688* repeats are relatively older than *Rsp-like* repeats, as it reveals deeply divergent clades separating extant *1.688* variants (Fig. 5). Several major clades contain repeats from cytobands spanning large physical distances across the X (*e.g.*, the basal clade contains repeats from cytobands 1, 3, 9, and 11 from all four study species (Figs. S14–15)), and together suggest a recurrent history in which a historical variant proliferated, spread across the X chromosome, and subsequently underwent local diversification. This diversification of *1.688* repeats largely pre-dated the speciation events that gave rise to the four species (Figs. 5, S14–15). We reach this conclusion upon finding repeated instances of cytoband-specific, well-supported clades comprised of repeats from all four species, with a branching pattern that matches the evolutionary history of the species (i.e., *D. melanogaster* repeats forming a clade sister to the repeats of the *simulans* clade species; Figs. S14–S15). The relative ages of *1.688* and *Rsp-like* repeats are further supported by our estimates of cluster age based on within-cluster repeat divergence (Figs. S18–S19, Table S2). Finally, our observation that *1.688* is a relatively old satellite is consistent with similar conclusions from previous studies [33, 34, 70].

Regarding the fourth finding, the *Rsp-like* all-species tree shows evidence of two major interlocus expansions of *Rsp-like* repeats which occur as clades containing hundreds of repeats separated by short branches. One interlocus expansion occurred in the ancestor of the *simulans* clade, hereafter called the ‘sim-clade’ expansion (Figs. 5 and S16); the second expansion occurred within *D. simulans* alone, hereafter called the ‘sim-specific’ expansion (Figs. 5 and S17). Repeats from *D. mauritiana* account for 58.9% (n=178) of terminals in the sim-clade expansion and include repeats from cytoband 1, 3, 5, 11, and 12. Repeats from cytobands 1 and 2 in *D. sechellia* make up 25.8% (n=78) of terminals in the sim-clade expansion. The remaining 15.2% (n=46) of repeats are from cytobands 2–5 in *D. simulans*. The sim-specific expansion comprises 226 *D. simulans* repeats from cytobands 1, 3, 7, 8, 9, 10, 11, and 12, all separated by extremely short branches (sim-specific expansion; Figs. 5 and S16); this expansion accounts for most of the increased *Rsp-like* repeats in *D. simulans* relative to all other species. The patterns in this all-species tree (Figs. 5 and S16–17) suggest that the *Rsp-like* repeats proximal to cytoband 6 in *D. simulans* and *D. mauritiana* arose through independent interlocus expansions of *Rsp-like* repeats. However, the effects of gene conversion could have erased evidence of a common origin of the expansions.

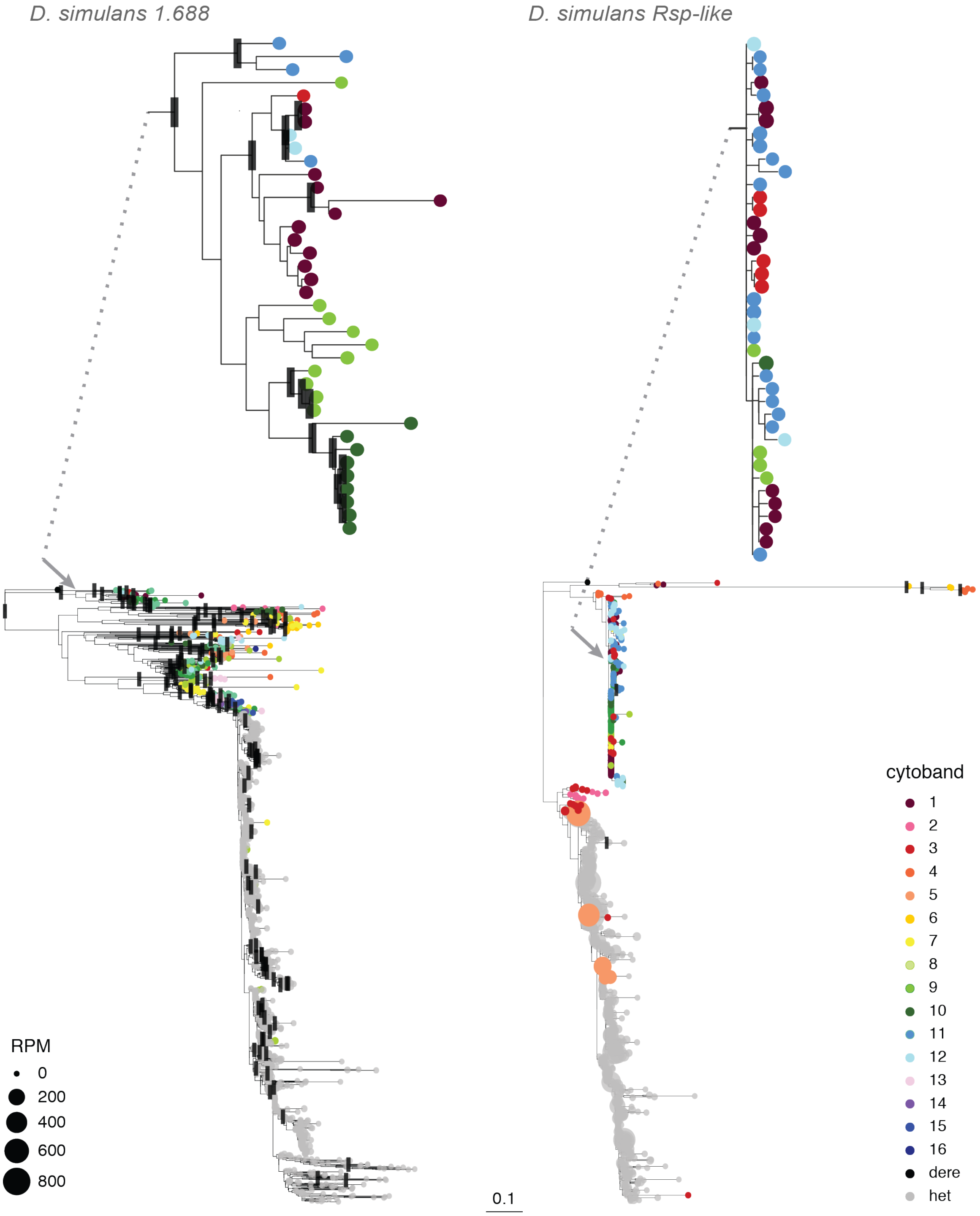


Figure 4. Comparison of phylogenetic patterns *1.688* and *Rsp-like* for *D. simulans*. Each terminal represents an individual repeat monomer from the X chromosome. Colored tip terminals indicate euchromatic repeats; gray tip terminals represent repeats from heterochromatic loci (defined as unassigned scaffolds in the assembly). Black rectangles indicate nodes with bootstrap support ≥ 90. Two regions in each tree are shown in greater detail to highlight differential phylogenetic patterns observed in euchromatic repeats of *1.688* and *Rsp-like*; arrows and dotted lines indicate relative position of

enlarged regions in the tree. Branch lengths shown are proportional to divergence with both trees shown on the same relative scale. Sizes of the tips are scaled to reflect proportion of eccDNA reads mapping to a given variant, expressed as reads-per-million (RPM) (see eccDNA analysis). Maximum likelihood trees were inferred in RAxML with nodal support calculated following 100 bootstrap replicates.

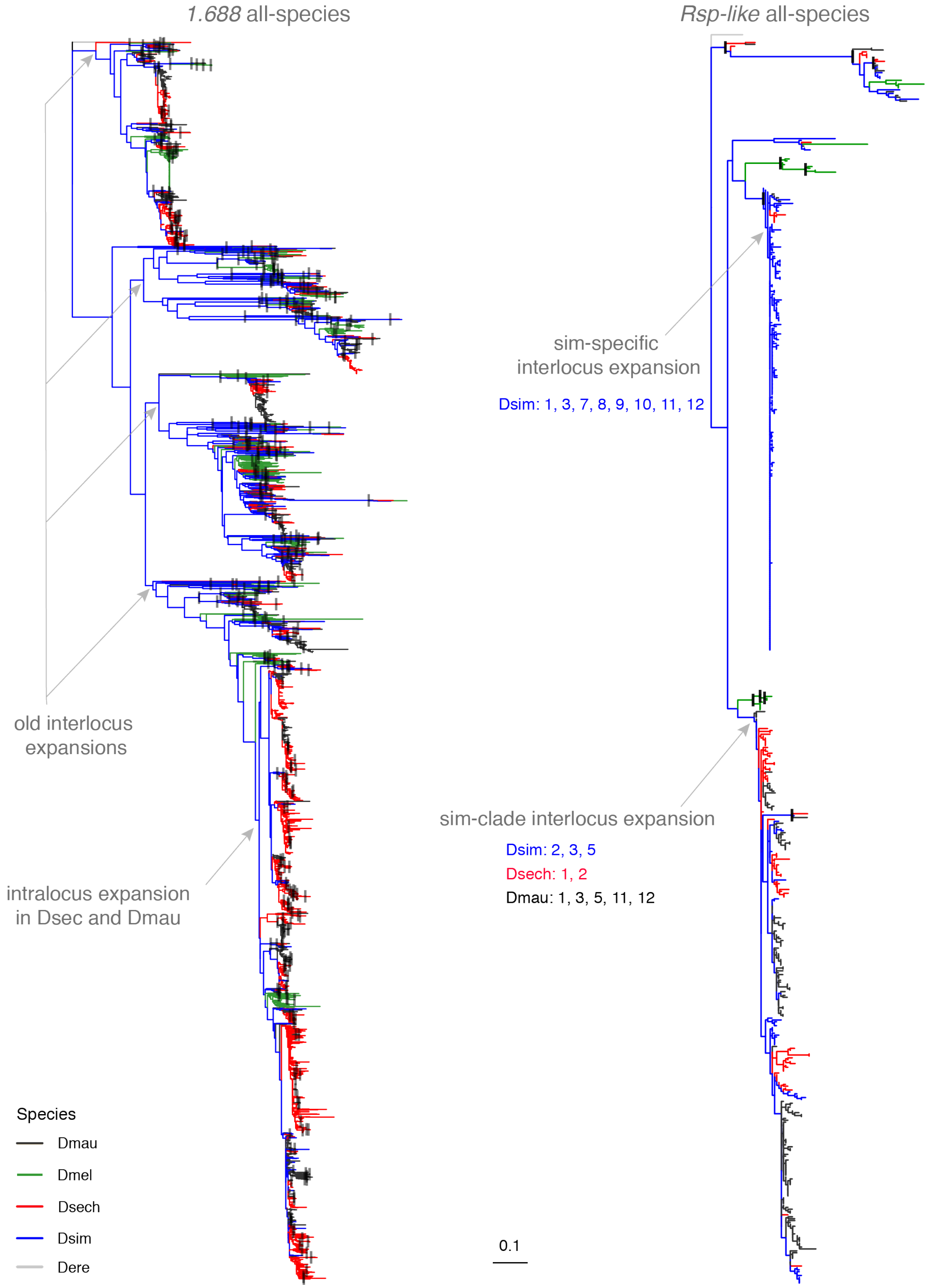

1.688 all-species
Rsp-like all-species
old interlocus expansions
intralocus expansion in Dsec and Dmau
sim-specific interlocus expansion
Dsim: 1, 3, 7, 8, 9, 10, 11, 12
sim-clade interlocus expansion
Dsim: 2, 3, 5
Dsech: 1, 2
Dmau: 1, 3, 5, 11, 12
Species
Dmau
Dmel
Dsech
Dsim
Dere
0.1

Figure 5: All-species maximum likelihood trees of euchromatic *1.688* and *Rsp-like*. Each terminal represents an individual repeat monomer. All monomers from clusters with

≥three repeats were included in the analysis. Species identity is indicated by branch color. Major inter and intralocus expansions of satellites discussed in the text are labeled with gray arrows. For interlocus expansions in *Rsp-like*, the species involved are listed along with cytological bands that are represented by monomers within the expansion. The outgroup (*D. erecta*) is indicated by gray branches. Black rectangles indicate nodes with bootstrap support ≥ 90. Maximum likelihood tree was inferred in RAxML with nodal support calculated following 100 bootstrap replicates. Branch length is shown proportional to relative divergence with both trees on the same relative scale. See Figures S14–17 for added detail as to genomic location of terminals.

Mechanisms driving satellite DNA turnover in the euchromatin

How did these new *Rsp-like* clusters arise? We found frequent co-localization of *Rsp-like* and *1.688* repeats in the two species with *Rsp-like* clusters at new genomic loci, which was surprising because these two repeats are unrelated at the sequence level. We therefore hypothesized that regions of microhomology could facilitate insertion of new *Rsp-like* repeats into pre-existing *1.688* clusters.

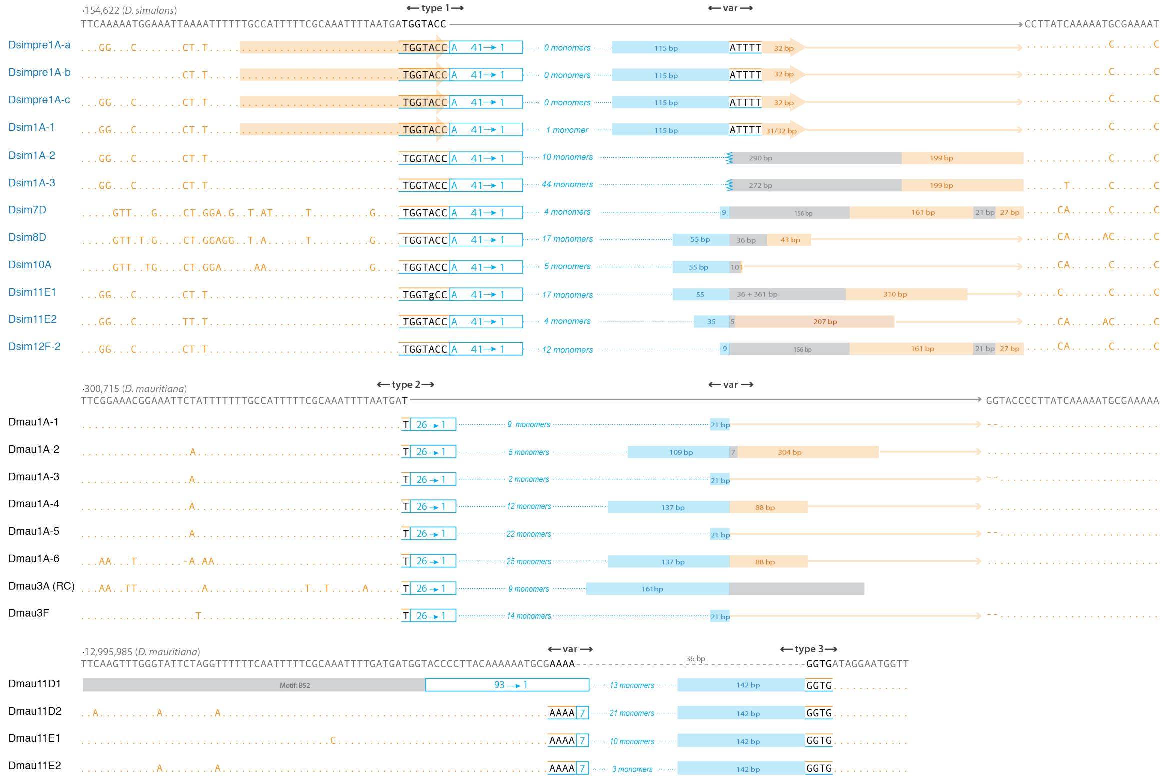
Our analysis of the *1.688*/*Rsp-like* junctions on each side of newly inserted *Rsp-like* clusters in *D. simulans* and *D. mauritiana* revealed multiple independent insertion events with shared signatures (Fig. 6). One prominent signature is that junctions between the *Rsp-like* and *1.688* sequences commonly occur at positions of microhomology that are shared between the satellites. The same junction sequence is often shared between clusters at different locations across the X chromosome. We use the sequence of these microhomologies to define clusters of the same "type": type 1 was found in *D. simulans* and types 2 and 3 were found in *D. mauritiana*. Because there are two different *1.688* variants adjacent to both type 1 and 2 junctions (*e.g.*, compare Dsim10A and Dsim11E1, Fig. 6), we infer that at least five independent events have created the three junction types.

Opposite the characteristic junctions, the other end of newly inserted *Rsp-like* clusters have more variable junctions within a type. For example, in *D. simulans*, type 1 is the predominant junction and is observed in 19/31 *Rsp-like* clusters located near *1.688* repeats, 12 of which are diagrammed in Figure 6. The type 1 junction is associated with a

42/49 bp truncated *Rsp-like* monomer abutting *1.688* sequences. The transition between the two satellite types includes a 7-bp region of microhomology ('TGGTACC'). Among these 12 *Rsp-like* clusters there are, however, at least 6 different junction sequences at the other end of the cluster. These include four clusters in which the sequences adjacent to *Rsp-like* are a duplication of the 32 bp (including the microhomology) of *1.688* sequences found at the type 1 junction. The remaining clusters have varying lengths of unannotated (5 bp to 290 bp) and *1.688* sequences (1 bp to 310 bp) in the variable junction region. A less obvious signature is that the new *Rsp-like* insertions, including clusters at 3F, 9D, 9F, 11C, 11D, 12C, and 12F-1 not diagrammed in Figure 6, are associated with a minor subset of *1.688* repeat variants. The two *1.688* variants found at the type 1 junction are only found in 121 (15.4%) of the 787 monomers in our alignment in *D. simulans*.

In *D. mauritiana*, type 2 clusters show a similar signature to *D. simulans* type 1 clusters: one side of the cluster shows a characteristic junction which is associated with a *Rsp-like* truncated monomer abutting *1.688* sequences, with the other side of the cluster showing more variable patterns. Interestingly, type 2 junctions occur at the same position within the *1.688* monomer and in a similar subset of variants as the *D. simulans* type 1 junction, however, the position in *Rsp-like* monomers associated with the junction differs between the two species (*i.e.* note 26/27 bp truncated monomers in *D. mauritiana* and 42/49 bp truncations in *D. simulans*) (Fig. 6). The variable side of the cluster shows 4 different sequences associated with the junction. The most common variable junction occurs in four of the eight clusters and has a 2 bp deletion before continuing with the interrupted *1.688* repeat sequence. Likewise, the 4 new clusters in cytoband 11 of *D. mauritiana* show these junction signatures although unlike the type 1 and type 2 junctions, these type 3 junctions have a deletion (36 bp) in the associated *1.688* sequences.

The particular double-strand break repair pathway(s), either non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ), used to insert the *Rsp-like* sequences is difficult to determine based solely on these observed sequences. NHEJ does not require, but can use, short stretches of microhomology (< 5 bp) [71]. The 4-8 bp of microhomology observed in the type 2 and type 3 junctions instead seem more consistent with those pathways employing MMEJ. The nature of the variable junctions (unannotated sequences/sequence variation in *1.688* repeat monomers) makes it difficult to determine whether microhomology is present. However, in two cases short runs of mononucleotides are present at the overlap between *1.688* and *Rsp-like* sequences. Together the patterns we observe are consistent with microhomology repair pathways giving rise to the new *1.688/Rsp-like* associations (Fig. 7a).



·154,622 (D. simulans)
←type 1→
←var→
TTCAAAAATGGAAATTAAAATTTTTTGCCATTTTTCGCAAATTTTAATGATGGTACC
CCTTATCAAAAATGCGAAAAT
Dsimpre1A-a TGGTACC A 41→1 0 monomers 115 bp ATTTT 32 bp
Dsimpre1A-b TGGTACC A 41→1 0 monomers 115 bp ATTTT 32 bp
Dsimpre1A-c TGGTACC A 41→1 0 monomers 115 bp ATTTT 32 bp
Dsim1A-1 TGGTACC A 41→1 1 monomer 115 bp ATTTT 31/32 bp
Dsim1A-2 TGGTACC A 41→1 10 monomers 290 bp 199 bp
Dsim1A-3 TGGTACC A 41→1 44 monomers 272 bp 199 bp
Dsim7D TGGTACC A 41→1 4 monomers 9 156 bp 161 bp 21 bp 27 bp
Dsim8D TGGTACC A 41→1 17 monomers 55 bp 36 bp 43 bp
Dsim10A TGGTACC A 41→1 5 monomers 55 bp 10
Dsim11E1 TGGTgCC A 41→1 17 monomers 55 36 + 361 bp 310 bp
Dsim11E2 TGGTACC A 41→1 4 monomers 35 5 207 bp
Dsim12F-2 TGGTACC A 41→1 12 monomers 9 156 bp 161 bp 21 bp 27 bp
·300,715 (D. mauritiana)
←type 2→
←var→
TTCGGAAACGGAAATTCTATTTTTTTGCCATTTTTCGCAAATTTTAATGAT
GGTACCCCTTATCAAAAATGCGAAAAA
Dmau1A-1 T 26→1 9 monomers 21 bp
Dmau1A-2 T 26→1 5 monomers 109 bp 7 304 bp
Dmau1A-3 T 26→1 2 monomers 21 bp
Dmau1A-4 T 26→1 12 monomers 137 bp 88 bp
Dmau1A-5 T 26→1 22 monomers 21 bp
Dmau1A-6 T 26→1 25 monomers 137 bp 88 bp
Dmau3A (RC) T 26→1 9 monomers 161bp
Dmau3F T 26→1 14 monomers 21 bp
·12,995,985 (D. mauritiana)
←var→
←type 3→
TTCAAGTTTGGGTATTCTAGGTTTTTTCAATTTTTCGCAAATTTTGATGATGGTACCCCTTACAAAAAATGCGAAAA
36 bp
GGTGATAGGAATGGTT
Dmau11D1 Motif: B52 93→1 13 monomers 142 bp GGTG
Dmau11D2 AAAA 7 21 monomers 142 bp GGTG
Dmau11E1 AAAA 7 10 monomers 142 bp GGTG
Dmau11E2 AAAA 7 3 monomers 142 bp GGTG

Figure 6. Junctions at new *Rsp-like* insertions in *D. simulans* and *D. mauritiana*. Junctions from a subset of the newer *Rsp-like* clusters (blue text/lines/boxes) are aligned and grouped into three types based on common signatures with nearby *1.688* monomers (orange text/lines/boxes). Type 1 is found in *D. simulans* while types 2 and 3 junctions are found in *D. mauritiana* (cytoband location of each cluster is indicated in the names at far left). Within each type, identical truncated *Rsp-like* monomers abut *1.688* at the same position in the *1.688* repeat monomer. In all three junction types, there is overlap between the two satellite sequences (black text) which, for at least the longer overlaps, potentially represents microhomology involved in the original insertion event. The second junction associated within and among these types is more variable ("var" in figure) with *Rsp-like* sequences abutting different positions of the *1.688* repeat or different unannotated sequences (gray boxes). The number of full length *Rsp-like* monomers as well as the lengths of truncated *Rsp-like* monomers, unannotated regions, and *1.688* sequences in this variable region are indicated for each cluster. Note that some clusters are nearly identical across this variable region (*e.g.,* Dsim7D and Dsim12F). The *1.688* sequences in the region that would be sequential to those sequences at the conserved junctions (dark gray

text above each junction type is the sequence within a specific *1.688* monomer) are indicated at the far right. Orange arrows in the first four *D. simulans* clusters indicate a duplication of the *1.688* sequences at the two junctions.

As described above, the relatively minor *1.688* repeat variants adjacent to the type 1 and type 2 junctions are each shared across multiple *1.688/Rsp-like* clusters (Fig. 6). This suggests either *Rsp-like* has repeatedly inserted into a particular subset of variants in both species, or that the multiple *1.688/Rsp-like* junctions were not formed independently within either species. In the latter scenario, a relatively rare microhomology-mediated event gives rise to a *1.688/Rsp-like* hybrid repeat, which then seeds new *Rsp-like* clusters at loci where *1.688* clusters were already present, facilitated by homology of the *1.688* portion of the novel hybrid repeat. We tested two predictions arising from this model: (1) newly inserted *Rsp-like* clusters would only occur at genomic loci where *1.688* repeats were already present; (2) any *1.688* sequences moving as a higher order repeat along with *Rsp-like* sequences may be differentiated from *1.688* sequences already present in clusters where the new insertions occurred, generating discordant phylogenetic relationships.

We tested the above predictions using *D. simulans Rsp-like* clusters with type 1 junctions (Fig. 6), focusing on the 12 of 19 clusters that are present at genomic loci where *Rsp-like* clusters are lacking in one or more of the other three study species (*i.e.*, those clusters at cytobands 7-12). We conducted a synteny analysis across species to establish orthology of the 12 clusters. If a *1.688* cluster was present at a syntenic position in the other species, we inferred that *Rsp-like* moved into an existing cluster in *D. simulans*. We found that all 12 new *Rsp-like* clusters in *D. simulans* had *1.688* repeats at that same location in each of the other three species with the exception of a single locus in *D. melanogaster* (Table S3). In *D. mauritiana,* all but two loci at cytoband 11 are missing *Rsp-like* repeats at the syntenic loci in all other species (Table S3). The fact that *1.688* clusters were already present at the site of new *Rsp-like* insertions suggests it is sequence homology (and/or microhomology) with *1.688* repeats that is facilitating new insertions. Testing for discordant phylogenetic relationships of *1.688* monomers surrounding *Rsp-like* junctions showed that in six of 12 clusters, the *1.688* repeat immediately adjacent to the *Rsp-like* junction shows strongly discordant relationship with the other *1.688* repeats in the cluster (Table S3), suggesting that at least a partial *1.688* repeat has moved together with *Rsp-like* repeats in the case of multiple new *Rsp-like* insertions.

Our findings from the *1.688*/*Rsp-like* junction and synteny analyses are consistent with a model in which small regions of microhomology can facilitate the integration of *Rsp-like* into *1.688*. Once this association is created, the larger regions of homology (*e.g.*, larger segments of flanking *1.688* repeats) may facilitate the rapid spread of *Rsp-like* across the chromosome (Fig. 7b,c), including through the movement of entire mixed clusters to new locations as a higher-order unit (Fig. 7d).

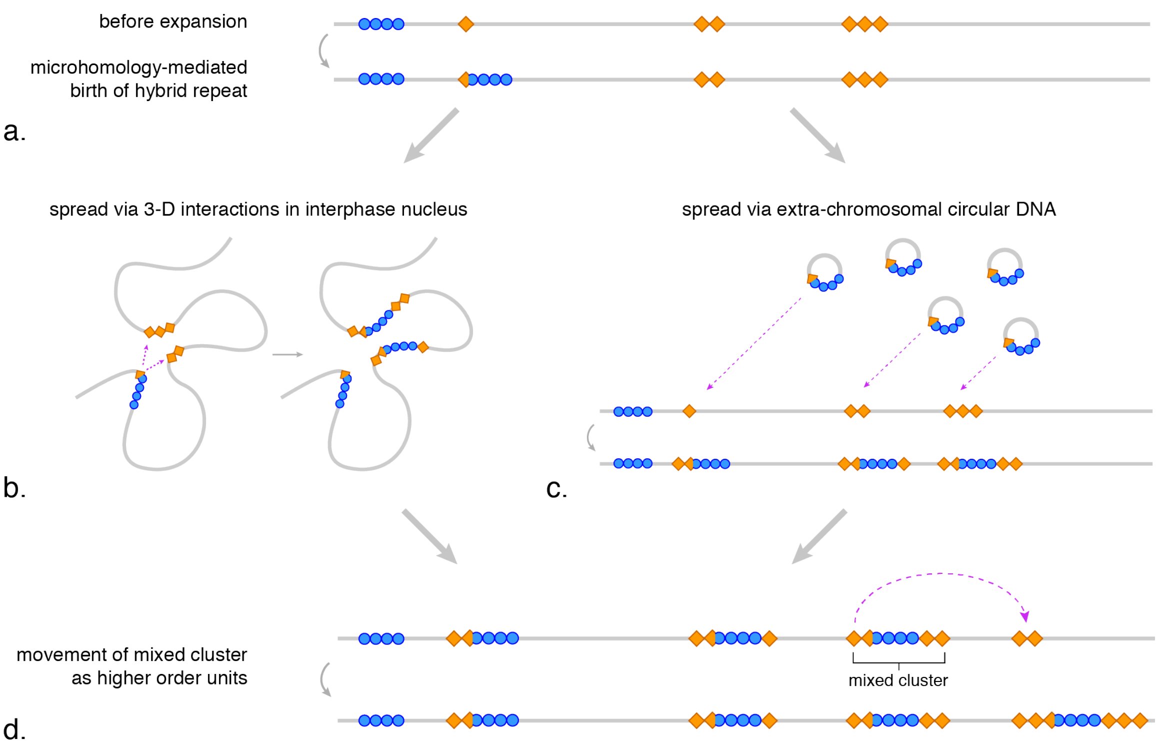

before expansion
microhomology-mediated birth of hybrid repeat
a.
spread via 3-D interactions in interphase nucleus
b.
spread via extra-chromosomal circular DNA
c.
movement of mixed cluster as higher order units
mixed cluster
d.

Figure 7. Proposed mechanisms of satDNA dynamics. Blue circles represent an ancestrally rare satellite (*i.e.*, *Rsp-like*), orange diamonds represent an abundant satellite present at many loci (*i.e.*, *1.688*), gray lines represent a fraction of a chromosome that spans many megabases. (a.) illustrates the microhomology-mediated birth of a hybrid repeat formed from the rare+common satellites*,* facilitating spreading of the rare satellite to loci where the abundant satellite is already present through processes illustrated by b–d. (b.) loci that are physically distant on a linear X chromosome may interact in three-dimensional space within the interphase nucleus, interlocus gene conversion of orange

satellite repeats may then facilitate the spread of blue repeats. (c.) satellite DNAs are present on extrachromosomal circular DNAs, which may facilitate their spread to new loci. (d.) after new insertions of the blue satellite, entire mixed clusters may move as higher order units. The mechanisms illustrated in (b) and (c) could also be responsible for the generation of the hybrid repeat (a) and movement of higher order units (d). Not illustrated is the expansion or contraction of a repeat cluster at a given locus due to unequal exchange with a different cluster of the same repeat type.

Mechanisms underlying spread of clusters to new loci

Two mechanisms that can explain the spread of nearly identical repeats across long physical distances are: (1) three dimensional interactions in the nucleus creating opportunities for interlocus gene conversion between repeats over long linear distances; and (2) the spread of repeats via extrachromosomal circular DNA (eccDNA) to new loci across the X chromosome (Fig.7). Either mechanism can potentially explain the generation of new clusters or the spread of higher order units to new loci. Our reanalysis of *D. melanogaster* Hi-C data [72] provides some evidence of inter-cytoband interactions, particularly across the middle of the X chromosome (*i.e.*, from cytobands 6 through 14) (Fig. S20).

If long-distance gene conversion is facilitated by 3D interactions in the nucleus, we might expect *1.688* repeats and neighboring *Rsp-like* repeats to show a similar pattern of gene conversion. The circle plot of genetic distance between *Rsp-like* clusters in *D. simulans* shows a high degree of similarity, as evidenced by the preponderance of blue lines connecting cytobands across the X chromosome (Fig. S21). However, the *1.688* repeats adjacent to these *Rsp-like* clusters showed a mixed pattern, with high sequence similarity among repeats only at cytobands 1, 11, and 12. The majority (64.5%) of *1.688* repeats have <95% sequence similarity with any repeat from another cytoband, while the nearest *Rsp-like* repeat shows >95% similarity with repeats from multiple different cytobands. Thus, we find limited evidence of long-distance gene conversion in *1.688* sequences; however, it is possible that the older age and smaller size of *1.688* clusters relative to *Rsp-like* clusters may limit interlocus gene conversion.

eccDNA as a mechanism of satDNA to new genomic loci

Spread of repeats via eccDNA (extrachromosomal circular DNA) is another non-mutually exclusive mechanism that could mediate the spread of *Rsp-like* satellite repeats. We used 2D gel analysis to confirm/show the presence of *1.688* [48] and *Rsp* eccDNA in *D. melanogaster* (Fig. S22) and then isolated (Fig. S22–23) and sequenced the eccDNA component from all four species. We estimated the abundance of sequences in eccDNA and in the genomic control using reads-per-million (RPM).

Not surprisingly, long-terminal repeats (LTRs) and complex satellites, including *1.688* and *Rsp-like*, are abundant on eccDNAs in all four species (Fig. S24; Fig. 8). In general, we find a strong correlation between the abundance of a repetitive element in the genome (estimated by RPM for that element in the non-digested gDNA Illumina control) and the abundance of eccDNA reads derived from that repeat. However, some repeats produce more eccDNA than expected given their genomic abundance (Fig. 8). *Rsp-like* repeats are particularly abundant on eccDNA in *D. simulans* (Fig. 8), where they comprise ~3% of the total eccDNA-enriched reads (a 24.5-fold enrichment over the undigested control), and in *D. sechellia* where they comprise ~4.9% of reads (a 5.75 enrichment over the undigested control).

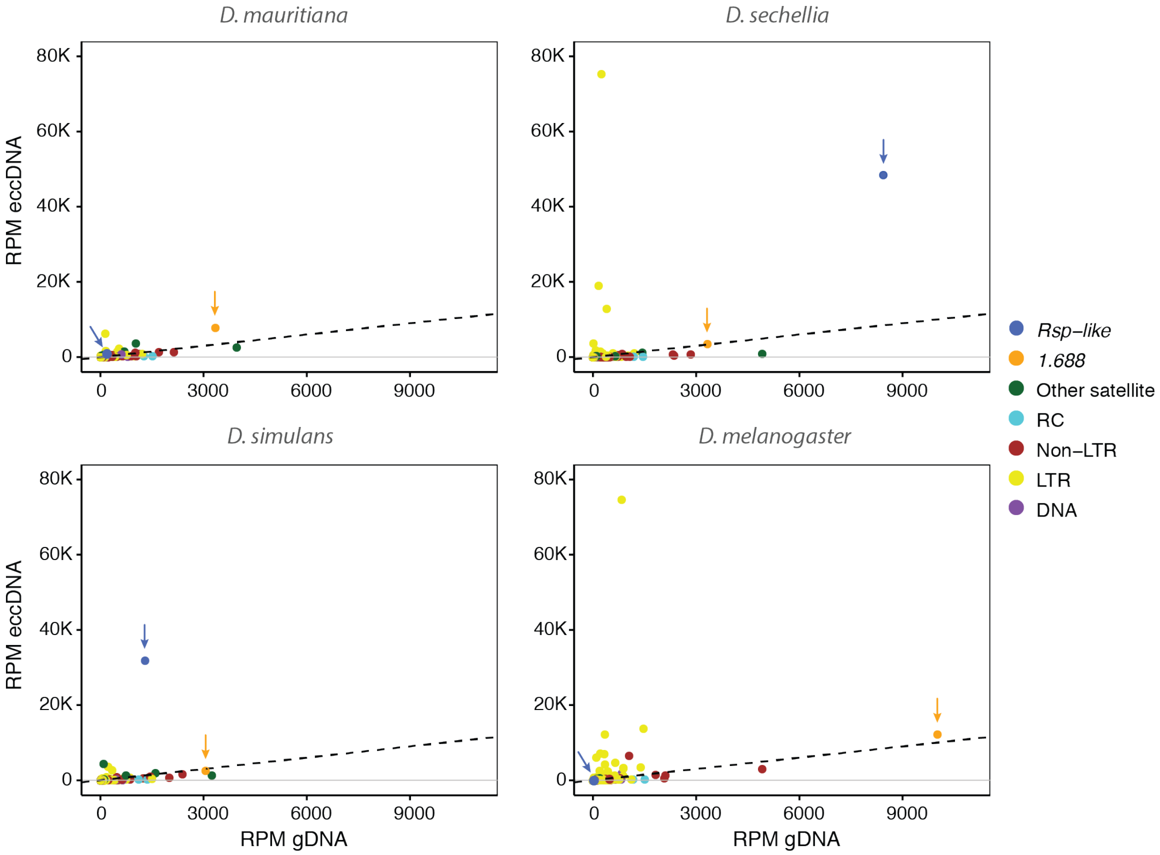


Figure 8. Scatter plot of eccDNA RPM and genomic DNA RPM. Repeats in the genome are categorized into Other satellite (complex satellites except *1.688* and *Rsp-like*), LTR retrotransposon, non-LTR retrotransposon, DNA transposon and rolling-circle (RC) transposon and are shown in different colors. *Rsp-like* (shown in blue) and *1.688* (shown in orange) are indicated by arrows. Dotted lines represent the same abundance of eccDNA and genomic DNA such that dots above the dotted line indicate repeats that are enriched in eccDNA libraries relative to genomic controls.

To determine the genomic source of satellite-derived eccDNAs, we estimated abundance of each sequence variant of *1.688* or *Rsp-like* from euchromatic and heterochromatic loci. We represent the estimated eccDNA abundance on phylogenetic trees by scaling tip labels based on the RPM of each variant (Figs. 4, S6–13). With the exception of *1.688* in *D. sechellia* and *D. mauritiana,* heterochromatic repeat variants produce more eccDNA. Consistent with the lack of heterochromatic *Rsp-like* repeats [58], few eccDNAs map to *D. mauritiana Rsp-like*. Some individual repeats generate more eccDNAs than others, possibly due to sequence composition, chromatin structure, and/or recombination environment. For example, in *D. simulans,* eight euchromatic *Rsp-like* variants from cytoband 5A are enriched for eccDNA (RPM ranges from ~100–600, see light orange tips on Figs. 4, S11). These euchromatic repeats group with the heterochromatic repeats that are also enriched for eccDNA reads (Figs. 4, S11). It is therefore possible that the repeats at 5A may be a result of a recent integration of heterochromatic-derived eccDNA carrying *Rsp-like* repeats.

DISCUSSION

We show that complex satDNAs have dynamic evolution over short evolutionary time scales with fine-scale resolution. Heterochromatic satDNA loci evolve rapidly in genomic location and abundance, consistent with previous studies [9, 60]. We show that euchromatic satellites are also fluid over short evolutionary timescales. Despite diverging from a common ancestor just 240K years ago [73], the *simulans* clade species differ in the total number of repeats, the number of clusters, and in the composition of clusters across syntenic loci (Figs. 1–3, Tables 2, S3). At least some of the differences in repeat abundance between species may be explained by ecology and demographic history. For example, *D. sechellia* is an island endemic with a historically low effective population size [68] and natural selection may be less efficacious in this species [67]. Interestingly, this species has larger euchromatic satDNA clusters suggesting that intralocus expansions of repeats may be weakly deleterious, but it does not have more discrete repeat loci. In contrast to *D. sechellia*, we see the proliferation of *Rsp-like* repeats in *D. simulans* and *D. mauritiana,* giving rise to new *Rsp-like* clusters across the X chromosome.

Our finding that X-linked euchromatic *1.688* has an old history of diversification is consistent with previous studies [33, 34]. Our detailed phylogenetic study of these repeats suggests an evolutionary history characterized by long periods of local differentiation among repeats, punctuated by the occasional proliferation of a particular variant, and subsequent local diversification (Figs. 5, S14-15). On a more recent time scale, our findings reveal that new *Rsp-like* clusters have spread across the X chromosomes of *D. simulans* and *D. mauritiana*, inserting into existing *1.688* clusters (Figs. 2, 4–5, S2, S7, S9, S11). Thus, our dissection of *Rsp-like* patterns in these species provides a glimpse into recent satellite proliferation dynamics that may implicate common processes underlying the evolution of both repeat types.

Mechanisms of Rsp-like *movement*

We find evidence that microhomology-mediated events generated a new hybrid repeat that joined the sequence of a relatively uncommon satellite (*i.e.*, *Rsp-like*) to that of an abundant satellite with a dense distribution across the X chromosome (*i.e.*, *1.688*). The birth of new *1.688*/*Rsp-like* hybrid repeats appears to have occurred independently in *D. simulans* and *D. mauritiana*, and likely multiple times within each species (Figs. 5–6, S16–17, Table S3). Microhomology-mediated repair events are implicated in creating structural rearrangements and chromosomal translocations across organisms (reviewed in [74]), as well as copy number variations associated with human disease [75], and gap repair after P-element transpositions in *Drosophila* [76, 77]. The original associations between *1.688* and *Rsp-like* repeats appear to be mediated by microhomology (*e.g.*, through MMEJ) in a single, or few independent events, however the larger regions of homology in the newly formed *1.688*/*Rsp-like* hybrid variant likely facilitated additional spread of *Rsp-like* clusters (Fig. 6, Table S3).

Mechanisms facilitating long-distance spread of new clusters

Questions remain about the source of the template *Rsp-like* sequences. We discussed two possibilities here: interlocus gene conversion and eccDNA reintegration. The complexity of the sequences observed in the *Rsp-like/1.688* variable junctions could also implicate pathways such as FoSteS (fork stalling and template switching, [78]) or MMBIR (microhomology-mediated break-induced replication, [75]). Both of these repair pathways occur during aberrant DNA replication and can involve multiple template switches facilitated by microhomology. The non-canonical termination of homologous recombination in mammalian cells resulting in complex breakpoints is also been linked to MMEJ/MMBIR [79](reviewed in [80]). During double-strand break (DSB) repair, synthesis-dependent strand annealing with an interlocus template switch may result in gene conversion events (*e.g.*, [81]) that insert *Rsp-like* sequences into existing *1.688* clusters. Similar events occur at the yeast MAT locus during gene conversion, where interchromosomal template switches occur even between divergent sequences, and these events can proceed based on microhomologies as small as 2 bp [82]. DNA prone to forming secondary structures (*e.g.,* non-B form DNA like hairpins or G quartets) can cause replication fork collapse that leads to DSB formation (reviewed in [83]). Blocks of complex satDNAs may be enriched for sequences that form secondary structures and therefore may have elevated rates of DSBs compared to single copy sequences. Elevated rates of DSB may make it more likely to observe non-homologous recombination-mediated repair events resulting in complex rearrangements, differences in repeat copy number and, as we describe here, the colonization of repeats at new genomic positions across large physical distances.

We add another mechanistic insight into satDNA dynamics by showing that complex satellites are abundant on eccDNA (Figs. 8, S22-24), consistent with other studies showing that repeats generate eccDNA [84]. While the abundance of most eccDNAs correlates with their genomic abundance, some repeats, such as *Rsp-like* in *D. simulans,* generate excess eccDNAs. The formation of eccDNA may depend on DNA sequence, organization (*e.g.,* repetitive versus unique), chromatin status, and possibly its higher order structure. It is possible that the high abundance of *Rsp-like* derived eccDNA suggests that this satellite is unstable at the chromatin level, or more prone to DSB. EccDNA formation exploits different methods of DNA damage repair, including HR (using solo LTRs [85]), MMEJ ([84, 86]), and NHEJ [87]. The repetitive nature of *1.688* and *Rsp-like* makes it difficult to examine junctions in the extrachromosomal circles themselves. We do find evidence suggesting that HR can give rise to *Rsp-like* circles, however. An eccDNA arising from an intrachromatid exchange event between repeats within the same array, followed by the reintegration of that eccDNA at a new genomic location, may generate new arrays where the first and last repeat are truncated, but together would form a complete monomer. We see this pattern in four of the new *Rsp-like* arrays in *D. simulans* (Dsimpre1A-a, Dsimpre1A-b, Dsimpre1A-c, Dsim1A-1; Fig. 6) and two arrays in *D. mauritiana* (Dmau1A-4, Dmau1A-6; Fig. 6). It is thus conceivable that eccDNAs are involved in the generation of new *Rsp-like* clusters. EccDNAs may be a source of genomic plasticity within species [88]; we suspect that they also played a role in the proliferation of satDNAs in the *simulans* clade, thus contributing to X-linked repeat divergence between these species. Experimental approaches will help explicitly test the hypothesis that satDNA-derived eccDNAs reintegrate in the genome.

Functional impact of rapid evolution of satDNA

A growing body of research suggests that shifts in satellite abundance and location may have consequences for genome evolution. Large scale rearrangements or divergence in heterochromatic satDNA may lead to hybrid incompatibilities. In *D. melanogaster* a heterochromatic block of *1.688* satellite (359-bp) is associated with embryonic lethality in *D. melanogaster – D. simulans* hybrids [16, 89] through mechanisms that we do not yet understand. However, even variation in small euchromatic satDNAs can have measurable effects on gene regulation and thus may be important for genome evolution. Short tandem repeats in vertebrate genomes can affect gene regulation by acting as binding sites for transcription factors [90, 91]. Additionally, repeats can have an impact on local chromatin, which may affect nearby gene expression (*e.g.,* [38]). Novel TE insertions can cause small RNA-mediated changes in chromatin (*e.g.,* H3K9me2) that can spread to nearby regions and alter local gene expression [92]. In *D. melanogaster*, siRNA mediated chromatin modifications at some *1.688* repeats play a role in X chromosome recognition during dosage compensation [40-42]. The turnover in repeat composition in *D. simulans* and *D. mauritiana* at loci (*e.g.,* Fig. 3) with demonstrated effects on chromatin and MSL recruitment [41, 42] raises the possibility that dynamic evolution of euchromatic satDNAs may have functional consequences for dosage compensation.

CONCLUSIONS

SatDNA evolution is highly dynamic over short evolutionary time periods, where the composition of heterochromatin shifts between even closely related species. Similar to the heterochromatin, we observe that satDNA in the euchromatin is dynamic with repeats changing in abundance, location, and composition between closely related species. Our detailed study of euchromatic repeats revealed the proliferation of a rare satellite (*Rsp-*

like) across the X chromosome. *Rsp-like* spread by inserting into existing clusters of the older, more abundant *1.688* satellite (see schematic in Fig. 7). Intralocus satDNA expansions via unequal exchange and the movement of higher-order repeats further contribute to the fluidity of the repeat landscape. Our analysis suggests that euchromatic satDNA repeats experience cycles of repeat proliferation and diversification: the phylogenetic patterns we see in the much older *1.688* satDNA in these species suggests a similar, albeit older, history of repeat interlocus expansions and subsequent diversification. SatDNA proliferation in genomes is analogous to bursts of TE proliferation, however, satDNAs do not encode proteins that facilitate their spread. Instead satDNAs appear to largely spread through recombination mechanisms. Our study lays the foundation for further mechanistic studies of satDNA proliferation and the possible functional and evolutionary consequences of these dynamics.

MATERIALS AND METHODS

We aimed to characterize patterns and mechanisms underlying the evolution of two complex satellite DNAs, *1.688* and *Rsp-like*, over short evolutionary time scales in *Drosophila melanogaster* and the closely related species in the *simulans* clade *D. mauritiana*, *D. sechellia*, and *D. simulans*. We studied broad-scale patterns using classical cytogenetic and molecular biology techniques. We leveraged high-quality PacBio assemblies to characterize the dynamics of these repeats at base-pair resolution across the X chromosome. We tested hypotheses as to the mechanism mediating the insertion and spread to new genomic loci of expanding *Rsp-like* repeats in *D. simulans* and *D. mauritiana*, and explored the potential role of interlocus gene conversion within the nucleus and the potential role of eccDNA in facilitating the spread of expanding satellites across long physical distances on the X chromosome. Our methods are described in more detail here and in the Supplemental Information.

Fluorescence in-situ hybridization

We studied broad-scale dynamics of complex satellites by mapping the location of *1.688* and *Rsp-like* repeats on *Drosophila* chromosomes using FISH protocols outlined in Larracuente and Ferree (2015). Briefly, larval brains were dissected in 1× PBS, treated with a hypotonic solution (0.5% sodium citrate) and fixed in 1.8% paraformaldehyde, 45% acetic acid, and dehydrated in ethanol. For salivary glands, the same procedure was followed except for the treatment with hypotonic solution. We generated biotin- and digoxigenin-labeled probes using nick translation on gel-extracted PCR products from 360-bp (*D. simulans* DNA; 360F:'ACTCCTTCTTGCTCTCTGACCA'; 360R:'CATTTTGTACTCCTTACAACCAATACTA') [16], and *Rsp-like* (*D. sechellia* DNA; Rsp-likeF:'ACTGATTATCATCGCCTGGT'; Rsp-likeR:'GTAACTCCAGTTCGCCTGGT) [58]. For the *D. melanogaster* 1.688 probe, we generated biotin-labeled probes using nick translation on gel-extracted PCR products from 260-bp repeats (260F: 5′-TGGAAATTTAATTACGAGCT-3′; 260R: 5′-ATGAAACTGTGTTCAACAAT-3′) [56], which cross hybridize with all heterochromatic 1.688 repeats [93]. We made the simulans clade *1.688* probe in the same way (360F 5'-ACTCCTTCTTGCTCTCTGACCA-3', 360R 5'-CATTTTGTACTCCTTACAACCAATACTA-3' [16]). Probes were hybridized overnight at 30°C, washed in 4× SSCT and 0.1×SSC, blocked in a BSA solution, and treated with 1:100 Rhodamine-avadin (Roche) and 1:100 anti-dig fluorescein (Roche), with final washes in 4× SSCT and 0.1× SSC. Slides were mounted in Vecta-Shield with DAPI (Vector Laboratories), visualized on a Leica DM5500 upright fluorescence microscope at 100×, imaged with a Hamamatsu Orca R2 CCD camera, and analyzed using Leica's LAX software.

Repeat annotation

Repeat annotations were performed as described in [55]. Briefly, we constructed a custom repeat library by downloading the latest repetitive element release for *Drosophila* from RepBase and added custom satellite annotations. We manually checked our library for redundancies and miscategorizations. We used our custom library with RepeatMasker version 4.0.5 using permissive parameters to annotate the assemblies. We merged our repeat annotations with gene annotations constructed in Maker version 2.31.9 (for the *simulans* clade species) [94] or downloaded from Flybase (for *D. melanogaster*) [95].

We used custom Perl scripts to define clusters of satellites on the X chromosome and to determine the closest neighboring annotations. We defined clusters as two or more monomers of a given satellite within 500 bp of each other, though some analyses we also included single monomers. We grouped clusters according to cytoband (FlyBase annotation v6.03; ftp://ftp.flybase.net/releases/FB2014_06/precomputed_files/map_conversion/). We used custom scripts to translate the coordinates of cytoband boundaries from *Drosophila melanogaster* to the other three species with the following workflow. We extracted 30K bases adjacent to the coordinate of each cytoband sub-division in the *D. melanogaster* assembly and used that sequence as a query in a BLAST search against repeat-masked versions of the *simulans* clade species genomes. To obtain rough boundaries of *D. melanogaster* cytobands in each *simulans* clade species, we defined the proximal-most boundary as the proximal coordinate of the first hit (>1 kb in length) from each cytoband region. We defined the distal boundary arbitrarily as one base less than the proximal coordinate of the next cytoband.

Evolutionary relationship of satDNAs within and among species

We compared evolutionary histories of *1.688* and *Rsp-like* by generating phylogenetic trees for each repeat within species (referred to in the text as 'within-species' trees). We compared patterns across the resulting trees by focusing on four aspects of the topologies: (1) general patterns of nodal support and branch lengths; (2) the relationship of repeats in euchromatic vs heterochromatic genome regions; (3) the fraction of highly-supported clades for which all descendants are repeats from the same cytoband, or two adjacent cytobands; this comparison is intended to test the null hypothesis that repeats from physically nearby location (*e.g.*, those within a cluster, or from nearby clusters) are expected to homogenize via gene conversion and show greater sequence similarity than physically distant clusters – deviation from this null model (*i.e.,* a tree with a low fraction of repeats showing local homogenization) could indicate recent spread of repeats to new loci where local homogenization and subsequent differentiation from other clusters has not yet had time to accumulate; (4) the presence of clades containing repeats from cytobands that are physically distant (*i.e.*, non-adjacent) relative to the linear organization of the X chromosome, which could indicate historical exchange events that span large physical distances relative to the linear organization of the X chromosome.

In addition to the above within-species trees, we generated 'all-species' trees for both *1.688* and *Rsp-like* repeats, which combined monomers from all four species in the same analysis. The all-species allowed additional insight as to the relative timing of diversification within each satellite type. In addition, the *Rsp-like* all-species tree allowed us to test whether interlocus expansions of *Rsp-like* repeats in the *simulans* clade species share a common origin, or occurred independently. For all analyses, we aligned repeats using MAFFT [96] in Geneious v8.1.6 with the "auto" option, which selects the most efficient algorithm based on the number of input sequences. Prior to alignment, we filtered the data to exclude monomers originating in small clusters (*i.e.*, those with ≤ two monomers) and monomers below a minimum length (*i.e.*, ≤ 100bp for *Rsp-like*, ≤300bp for *1.688*). As outgroup sequences, we used consensus sequences of *Rsp-like* and *1.688* repeats from *Drosophila erecta*, a near relative of the study species. We used RaxML v8.2.11 to infer maximum likelihood trees with GTR+gamma as the model of evolution, and conducted bootstrap analysis using the --autoMRE option to automatically determine the optimal number of bootstrap replicates [97] which is recommended for large data sets in the program documentation. The resulting trees were plotted and stylized using APE and ggtree in R [98] and Adobe Illustrator.

Cluster age estimation

We analyzed differential patterns of gene conversion within a repeat array to estimate the relative age of a given cluster. This analysis was designed to test our conclusion that *Rsp-like* clusters in *D. simulans* and *D. mauritiana* at novel loci are due to new insertions against the alternative that these are actually older clusters that were lost in the other species, which are being maintained homogeneous by long-distance gene conversion. According to the accretion model of repeat evolution [99], repeats at the edges of a cluster should undergo gene conversion less often due to adjacent non-homologous sequence, which will cause them to become more diverged from sequences in the center of the cluster as mutations accumulate. Thus, we use the pattern of sequence divergence within a cluster to infer the age of that cluster. We expect older clusters to have low divergence between repeats within the center of the cluster and high divergence between the first/last repeats and the center of the cluster. A "new" cluster would not have had time to accrue mutations or homogenize its center repeats through gene conversion, so there should be less difference between the divergence between the first/last repeats to the center and the divergence within the center repeats. We define a metric, dY, as the maximum of two comparisons: (1) the first-center distance vs the within-center distance; and (2) the last-center distance vs within-center distance. We provide more details on this metric and our workflow for estimating cluster age in Supplemental Materials.

Analysis of 1.688/Rsp-like *junctions*

We tested the hypothesis that short regions of microhomology could facilitate the insertion of *Rsp-like* repeats at new genomic loci using two complementary approaches: (1) through extensive visual examination of *1.688/Rsp-like* junctions in *D. simulans* and *D. mauritiana* in the context of multi-sequence alignments as well as the X chromosome assembly in Geneious v8.1.6. (2) we used MEME [100] to computationally detect motifs that are enriched at the edges of new *Rsp-like* clusters. Additional details are provided in Supplemental Methods.

Analysis of syntenic 1.688 *clusters with* Rsp-like *insertions in* D. simulans

We tested the prediction that new *Rsp-like* clusters would insert only at loci where *1.688* clusters were already present by extracting 5 kb of sequence immediately upstream and downstream of the loci containing a mixed *1.688/Rsp-like* cluster in *D. simulans*. We determined the orthologous position of these flanking sequences in the other three study species by using the flanks as BLAST query sequences which we searched against custom BLAST databases built from the assemblies of the other species. We accepted best hits as orthologous sequences only if they were reciprocal best hits when BLASTed back against the *D. simulans* genome assembly. We then navigated to the orthologous flanking sequences of each cluster to determine whether a 1.688 cluster was present at that locus in the three other study species.

We tested for discordant phylogenetic relationships among *1.688* repeats in clusters with new *Rsp-like* insertions in *D. simulans* by extracting *1.688* repeats surrounding the *Rsp-like* insertion and flagging those sequences in a phylogenetic analysis in which they were included with all *1.688* euchromatic repeats from *D. simulans*. We extracted flanking sequences, generated custom BLAST databases, conducted BLAST searches, and extracted relevant *1.688* monomers in Geneious v.8.1.9. For both of the above tests, we used as models those *Rsp-like* clusters that show the dominant junction signature in *D. simulans* (Fig. 6), with a focus on 12 clusters that are present at genomic loci where *Rsp-like* clusters are lacking in one or more of the other three study species (*i.e.*, those clusters at cytobands 7-12).

Testing for gene conversion at 1.688/Rsp-like *junctions*

To test whether *1.688* clusters near *Rsp-like* clusters show evidence of recent gene conversion, we created all-by-all distance matrices of *Rsp-like* repeats. In addition, we created a similar distance matrix of all *1.688* repeats that are within 100 bases of a *Rsp-like* cluster. We plotted each distance matrix as a circular plot (similar to genome synteny plots) using BioCircos v0.3.4 [101]. In the resulting plot each repeat is grouped by cytoband, and any repeats with genetic distances ≤ 0.05 have connecting lines drawn between their position on the circle. Both *1.688* and *Rsp-like* plots were made on the same cytoband scale. This allowed us to overlay the *Rsp-like* and *1.688* plots in order to compare their patterns of sequence divergence at adjacent positions. We only plotted clusters with more than two repeats.

Extrachromosomal circular DNA isolation

Genomic DNA was isolated from 20 five-day adult females (20-25 mg) from *D. melanogaster* (strain iso 1), *D. mauritiana*, (strain 12), *D. sechellia* (strain C), and *D. simulans* (strain XD1) using standard phenol-chloroform extractions. The DNAs were ethanol precipitated and resuspended in 10 mM Tris-EDTA, pH 8.0. The concentrations were determined by Qubit fluorometric quantification. 200 ng of each genomic DNA was subjected to exoV (New England Biolabs) digestion as described by [102]. In short, after digestion at 37° for 24 hours, the DNAs were incubated at 70° for 30 minutes. Additional buffer, ATP, and exoV were then added and the samples incubated at 37° for another 24 hours. The process was repeated for a total of 4, 24 hour incubations with exoV. The concentration of the remaining DNA was determined by Qubit.

Verification of eccDNA in exoV digestion

Aliquots of the undigested genomic DNAs were diluted to the comparable volumes of samples after exoV digestion. A dilution series was then made for PCR analysis of both the exoV digested and the undigested DNAs. Primers used included those for rp49 [5’-CAGCATACAGGCCCAAGATC-3’, 5’-CAGTAAACGCGGTTCTGCATG-3’], tRNA(lysine) [5’-CTAGCTCAGTCGGTAGAGCATGA-3’, 5’-CCAACGTGGGGCTCGAAC -3’], mitochondria COXI [5’-GATCAAACAAATAAAGGTATACG-3’, 5’-GTTCCATGTAAAGTAGCTAATC-3’], 5S [5’-GCCAACGACCATACCACG-3’, 5’-GTGGACGAGGCCAACAAC-3’], *Rsp* [5’-GGAAAATCACCCATTTTGATCGC-3’, 5’-CCGAATTCAAGTACCAGAC-3’], *Rsp-like* [5’-ACTGATTATCATCGCCTGGT-3’, 5’-GTAACTCCAGTTCGCCTGGT-3’], *1.688* [mel 5’-5’GTTTTGAGCAGCTAATTACC-3’, mel 5’TATTCTTACATCTATGTGACC-3’ [103] and sech 5’-ACTCCTTCTTGCTCTCTGACCA-3’, sech 5’-CATTTTGTACTCCTTACAACCAATACTA-3’].

2D gel analysis

Genomic DNA was isolated from the Raleigh 370 strain of *D. melanogaster* as described above. 10 ug of DNA was fractionated by electrophoresis as described [48]. The DNA was then depurinated, denatured, and neutralized before being transferred overnight in high salt (20 X SSC/ 1 M NH4Acetate) to a nylon membrane (Biodyne, ThermoScientific). DNA was UV crosslinked and hybridizations were done overnight at 55°C in North2South hybridization buffer (ThermoScientific). Biotinylated RNA probes were generated from *Rsp* or *1.688* PCR generated amplicons as described previously [93]. The hybridized membrane was processed as recommended for the Chemiluminescent Nucleic Acid Detection Module (ThermoScientific), and the signal recorded on a ChemiDoc XR+ (BioRad).

eccDNA sequencing

We prepared eccDNA-enriched samples and genomic DNA control samples for Illumina sequencing using a NEBNext FS DNA Ultra II Library Prep Kit (New England Biolabs). To control for bias associated with differential PCR amplification among libraries, we used results from an initial round of library preparation to understand variation in library yield between eccDNA isolates and control samples. Initial bioanalysis traces revealed over-amplification in our genomic controls and probable primer/adapter dimers in our eccDNA-enriched samples. To eliminate over-amplification, we halved the amount of input in our control samples and used protocol modifications outlined in [104] to reduce adapter dimer content and maximize yield of eccDNA-enriched samples. We generated final libraries using 2 ng of input for eccDNA-enriched samples and 1 ng of input for control samples, with 13 amplification cycles for all samples to minimize amplification bias and allow comparison between samples. Bioanalysis of resulting libraries showed clean traces for all samples (*e.g.*, no evidence of primer/adapter dimer peaks or over amplification). Libraries were pooled and sequenced on the same 150-base paired-end lane of an Illumina HiSeq 4000 by GENEWIZ laboratories (South Plainfield, NJ, USA). Reads from the control and enriched samples were evaluated using FastQC and trimmed using Trimgalore, then were mapped to the genome using Bowtie2 default parameters. For the repeat composition analysis (Figs. 8 and S23), we used a heterochromatin enriched assembly for *D. melanogaster* [105], which has more complete repeat information in heterochromatin regions. Based on our repeat annotations, we calculated the reads per million (RPM) for each repeat using a custom python script. We calculated relative abundance of eccDNA for each repeat in each species by normalizing to its own undigested genomic DNA control. We excluded simple tandem satellite repeats (monomers of 5-12 bp) following analysis because of Illumina read bias from library preparation. To estimate the linear DNA contamination in our eccDNA enriched library, we calculated the RPM values for all genes in the genome (excluding histone cluster and rDNA loci) using HTSeq-count [106], and we found that the mean and median of gene RPMs in eccDNA enriched libraries are ~5% - ~20% of that in undigested genomic DNA control libraries for all species, suggesting effective enrichment of eccDNA in our eccDNA libraries.

Hi-C analysis of 3D interactions in D. melanogaster embryo

We used a publicly available Hi-C dataset from stage 16 embryos (Gene Expression Omnibus accession number GSE103625) to test the 3D interactions among satellite repeats in *D. melanogaster* [72]. We mapped Hi-C raw sequence reads to the r6 reference genome, and processed the output with the HiC-Pro pipeline [107] to obtain contact matrix at 10kb resolution (default parameters). We summarized results from the contact matrix in R using the Biocircos v0.3.4 [101]. We plotted inter-cytoband interactions using a cutoff of normalized interaction counts > 40 in 10-kb windows and excluded the *1.688* sequences themselves to avoid potential mappability issues (see supplemental materials).

DECLARATIONS

Funding

This work is funded by National Institutes of Health General Medical Sciences grant (R35GM119515), the University of Rochester, and a Stephen Biggar and Elisabeth Asaro Fellowship in Data Science. JS is supported by an NSF Postdoctoral Fellowship (DBI-1811930).

Acknowledgements

We thank Massa Shoura (Stanford University) for advice on eccDNA isolation. We also thank Tom Eickbush, Jack Werren, and Ching-Ho Chang for helpful feedback on the study.

Availability of data and materials

Illumina genomic DNA and eccDNA raw reads for each species will be available in NCBI's SRA (accession is forthcoming). All data files and code for analysis and producing plots are deposited in Github (https://github.com/LarracuenteLab/simulans_clade_satDNA_evolution) and in the Dryad Digital Repository (doi is forthcoming).

Competing interests

The authors declare that they have no competing interests.

AUTHOR INFORMATION

Author Affiliations:

University of Rochester, Department of Biology, 337 Hutchison Hall, Rochester, NY, 14627

Department of Biomedical Genetics, University of Rochester Medical Center, 601 Elmwood Ave. Rochester, NY, 14642

Current affiliations:

DEK: Harvard University

IW: Harvard University Massachusetts General Hospital

Author contributions

AML and DEK conceived the study, JSS and DGE helped further develop aspects of the study design. DEK, JSS, DGE, XW, SN, and IW conducted analyses; AML, JSS, DEK, and DGE interpreted the data. JSS, AML, DGE wrote the paper; All authors read and approved the final manuscript.

Consent for publication

All authors have read the manuscript and give their consent to submit.

REFERENCES

1. Kit S: **Equilibrium Sedimentation in Density Gradients of DNA Preparations from Animal Tissues.** *Journal of Molecular Biology* 1961, **3:**711-&.
2. Sueoka N: **Variation and Heterogeneity of Base Composition of Deoxyribonucleic Acids - a Compilation of Old and New Data.** *Journal of Molecular Biology* 1961, **3:**31-&.
3. Szybalski W: **Use of cesium sulfate for equilibrium density gradient centrifugation.** *Methods Enzymol* 1968, **12B:**330-360.
4. Britten RJ, Kohne DE: **Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms.** *Science* 1968, **161:**529-540.
5. Yunis JJ, Yasmineh WG: **Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation.** *Science* 1971, **174:**1200-1209.
6. Charlesworth B, Langley CH, Stephan W: **The evolution of restricted recombination and the accumulation of repeated DNA sequences.** *Genetics* 1986, **112:**947-962.
7. Charlesworth B, Sniegowski P, Stephan W: **The evolutionary dynamics of repetitive DNA in eukaryotes.** *Nature* 1994, **371:**215-220.
8. Ugarkovic D, Plohl M: **Variation in satellite DNA profiles--causes and effects.** *The EMBO journal* 2002, **21:**5955-5959.
9. Strachan T, Coen E, Webb D, Dover G: **Modes and rates of change of complex DNA families of Drosophila.** *J Mol Biol* 1982, **158:**37-54.
10. Blattes R, Monod C, Susbielle G, Cuvier O, Wu JH, Hsieh TS, Laemmli UK, Kas E: **Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide.** *EMBO J* 2006, **25:**2397-2408.
11. Dernburg AF, Sedat JW, Hawley RS: **Direct evidence of a role for heterochromatin in meiotic chromosome segregation.** *Cell* 1996, **86:**135-146.
12. Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA: **Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles.** *Genome Res* 2016, **26:**1301-1311.
13. Fishman L, Saunders A: **Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers.** *Science* 2008, **322:**1559-1562.
14. Fishman L, Willis JH: **A novel meiotic drive locus almost completely distorts segregation in mimulus (monkeyflower) hybrids.** *Genetics* 2005, **169:**347-353.
15. Lindholm AK, Dyer KA, Firman RC, Fishman L, Forstmeier W, Holman L, Johannesson H, Knief U, Kokko H, Larracuente AM, et al: **The Ecology and Evolutionary Dynamics of Meiotic Drive.** *Trends Ecol Evol* 2016, **31:**315-326.
16. Ferree PM, Barbash DA: **Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila.** *PLoS biology* 2009, **7:**e1000234.
17. Bosco G, Campbell P, Leiva-Neto JT, Markow TA: **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.

18. Hartl DL: **Molecular melodies in high and low C.** *Nat Rev Genet* 2000, **1:**145-149.
19. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Molecular biology and evolution* 1987, **4:**203-221.
20. Schlotterer C, Tautz D: **Slippage synthesis of simple sequence DNA.** *Nucleic Acids Res* 1992, **20:**211-215.
21. Southern EM: **Base sequence and evolution of guinea-pig alpha-satellite DNA.** *Nature* 1970, **227:**794-798.
22. Lohe AR, Brutlag DL: **Identical satellite DNA sequences in sibling species of Drosophila.** *J Mol Biol* 1987, **194:**161-170.
23. Walsh JB: **Persistence of tandem arrays: implications for satellite and simple-sequence DNAs.** *Genetics* 1987, **115:**553-567.
24. Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GC: **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.
25. McGurk MP, Barbash DA: **Double insertion of transposable elements provides a substrate for the evolution of satellite DNA.** *Genome Res* 2018, **28:**714-725.
26. Vondrak T, Avila Robledillo L, Novak P, Koblizkova A, Neumann P, Macas J: **Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats.** *Plant J* 2019.
27. Fry K, Salser W: **Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents.** *Cell* 1977, **12:**1069-1084.
28. Smith GP: **Evolution of repeated DNA sequences by unequal crossover.** *Science* 1976, **191:**528-535.
29. Schlotterer C, Tautz D: **Chromosomal homogeneity of Drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution.** *Curr Biol* 1994, **4:**777-783.
30. Dover G: **Molecular drive: a cohesive mode of species evolution.** *Nature* 1982, **299:**111-117.
31. Dover G: **Concerted evolution, molecular drive and natural selection.** *Current biology : CB* 1994, **4:**1165-1166.
32. Coen E, Strachan T, Dover G: **Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of Drosophila.** *J Mol Biol* 1982, **158:**17-35.
33. Waring GL, Pollack JC: **Cloning and characterization of a dispersed, multicopy, X chromosome sequence in Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 1987, **84:**2843-2847.
34. DiBartolomeis SM, Tartof KD, Jackson FR: **A superfamily of Drosophila satellite related (SR) DNA repeats restricted to the X chromosome euchromatin.** *Nucleic Acids Res* 1992, **20:**1113-1116.
35. Kuhn GC, Kuttler H, Moreira-Filho O, Heslop-Harrison JS: **The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes.** *Molecular biology and evolution* 2012, **29:**7-11.

36. King DG, Soller M, Kashi Y: **Evolutionary tuning knobs.** *Endeavour* 1997, **21:**36-40.

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

38. Feliciello I, Akrap I, Ugarkovic D: **Satellite DNA Modulates Gene Expression in the Beetle Tribolium castaneum after Heat Stress.** *PLoS Genet* 2015, **11:**e1005466.

39. Lucchesi JC, Kuroda MI: **Dosage compensation in Drosophila.** *Cold Spring Harb Perspect Biol* 2015, **7**.

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

41. Joshi SS, Meller VH: **Satellite Repeats Identify X Chromatin for Dosage Compensation in Drosophila melanogaster Males.** *Curr Biol* 2017, **27:**1393-1402 e1392.

42. Deshpande N, Meller VH: **Chromatin That Guides Dosage Compensation Is Modulated by the siRNA Pathway in Drosophila melanogaster.** *Genetics* 2018, **209:**1085-1097.

43. Lundberg LE, Kim M, Johansson AM, Faucillion ML, Josupeit R, Larsson J: **Targeting of Painting of fourth to roX1 and roX2 proximal sites suggests evolutionary links between dosage compensation and the regulation of the fourth chromosome in Drosophila melanogaster.** *G3 (Bethesda)* 2013, **3:**1325-1334.

44. Kim M, Ekhteraei-Tousi S, Lewerentz J, Larsson J: **The X-linked 1.688 Satellite in Drosophila melanogaster Promotes Specific Targeting by Painting of Fourth.** *Genetics* 2018, **208:**623-632.

45. Lieber MR, Yu K, Raghavan SC: **Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations.** *DNA Repair (Amst)* 2006, **5:**1234-1245.

46. Richardson C, Jasin M: **Frequent chromosomal translocations induced by DNA double-strand breaks.** *Nature* 2000, **405:**697-700.

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

48. Cohen S, Yacobi K, Segal D: **Extrachromosomal circular DNA of tandemly repeated genomic sequences in Drosophila.** *Genome Res* 2003, **13:**1133-1145.

49. Zellinger B, Riha K: **Composition of plant telomeres.** *Biochim Biophys Acta* 2007, **1769:**399-409.

50. Cohen Z, Bacharach E, Lavi S: **Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway.** *Oncogene* 2006, **25:**4515-4524.

51. Cohen S, Menut S, Mechali M: **Regulated formation of extrachromosomal circular DNA molecules during development in Xenopus laevis.** *Mol Cell Biol* 1999, **19:**6682-6689.

52. Paulsen T, Kumar P, Koseoglu MM, Dutta A: **Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells.** *Trends Genet* 2018, **34:**270-278.
53. Navratilova A, Koblizkova A, Macas J: **Survey of extrachromosomal circular DNA derived from plant satellite repeats.** *BMC Plant Biol* 2008, **8:**90.
54. Gallach M: **Recurrent turnover of chromosome-specific satellites in Drosophila.** *Genome Biol Evol* 2014, **6:**1279-1286.
55. Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion J, Montooth K, Meiklejohn C, Liao Y, Larracuente AM, Emerson JJ: **Evolution of genome structure in the Drosophila simulans complex species.** *in prep*.
56. Abad JP, Agudo M, Molina I, Losada A, Ripoll P, Villasante A: **Pericentromeric regions containing 1.688 satellite DNA sequences show anti-kinetochore antibody staining in prometaphase chromosomes of Drosophila melanogaster.** *Mol Gen Genet* 2000, **264:**371-377.
57. Losada A, Villasante A: **Autosomal location of a new subtype of 1.688 satellite DNA of Drosophila melanogaster.** *Chromosome Res* 1996, **4:**372-383.
58. Larracuente AM: **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* 2014, **14:**233.
59. Lohe AR, Roberts PA: **Evolution of satellite DNA sequences in *Drosophila*.** In *Heterochromatin: Molecular and Structural Aspects.* Edited by Verma RS. Cambridge, UK: Cambridge University Press; 1988
60. Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM: **Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex.** *G3 (Bethesda)* 2017, **7:**693-704.
61. Pimpinelli S, Dimitri P: **Cytogenetic analysis of segregation distortion in Drosophila melanogaster: the cytological organization of the Responder (Rsp) locus.** *Genetics* 1989, **121:**765-772.
62. Wu CI, Lyttle TW, Wu ML, Lin GF: **Association between a satellite DNA sequence and the Responder of Segregation Distorter in D. melanogaster.** *Cell* 1988, **54:**179-189.
63. Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS: **Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the Drosophila buzzatii cluster.** *Chromosome Res* 2008, **16:**307-324.
64. Plohl M, Petrovic V, Luchetti A, Ricci A, Satovic E, Passamonti M, Mantovani B: **Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks.** *Heredity (Edinb)* 2010, **104:**543-551.
65. Bigot Y, Hamelin MH, Periquet G: **Heterochromatin condensation and evolution of unique satellite-DNA families in two parasitic wasp species: Diadromus pulchellus and Eupelmus vuilleti (Hymenoptera).** *Mol Biol Evol* 1990, **7:**351-364.
66. Macgregor HC, Sessions SK: **The biological significance of variation in satellite DNA and heterochromatin in newts of the genus Triturus: an**

evolutionary perspective. *Philos Trans R Soc Lond B Biol Sci* 1986, **312:**243-259.

67. McBride CS: **Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia.** *Proc Natl Acad Sci U S A* 2007, **104:**4996-5001.
68. Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou ML: **Species-wide genetic variation and demographic history of Drosophila sechellia, a species lacking population structure.** *Genetics* 2009, **182:**1197-1206.
69. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B: **On the role of unequal exchange in the containment of transposable element copy number.** *Genet Res* 1988, **52:**223-235.
70. Hsieh T-S, Brutlag D: **Sequence and sequence variation within the 1.688 g/cm3 satellite DNA of Drosophila melanogaster.** *Journal of Molecular Biology* 1979, **135:**465-481.
71. Chang HHY, Pannunzio NR, Adachi N, Lieber MR: **Non-homologous DNA end joining and alternative pathways to double-strand break repair.** *Nat Rev Mol Cell Biol* 2017, **18:**495-506.
72. Ogiyama Y, Schuettengruber B, Papadopoulos GL, Chang JM, Cavalli G: **Polycomb-Dependent Chromatin Looping Contributes to Gene Silencing during Drosophila Development.** *Mol Cell* 2018, **71:**73-88 e75.
73. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC: **Genome sequencing reveals complex speciation in the Drosophila simulans clade.** *Genome research* 2012, **22:**1499-1511.
74. McVey M, Lee SE: **MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings.** *Trends Genet* 2008, **24:**529-538.
75. Hastings PJ, Ira G, Lupski JR: **A microhomology-mediated break-induced replication model for the origin of human copy number variation.** *PLoS Genet* 2009, **5:**e1000327.
76. McVey M, Adams M, Staeva-Vieira E, Sekelsky JJ: **Evidence for multiple cycles of strand invasion during repair of double-strand gaps in Drosophila.** *Genetics* 2004, **167:**699-705.
77. Adams MD, McVey M, Sekelsky JJ: **Drosophila BLM in double-strand break repair by synthesis-dependent strand annealing.** *Science* 2003, **299:**265-267.
78. Lee JA, Carvalho CM, Lupski JR: **A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.** *Cell* 2007, **131:**1235-1247.
79. Hartlerode AJ, Willis NA, Rajendran A, Manis JP, Scully R: **Complex Breakpoints and Template Switching Associated with Non-canonical Termination of Homologous Recombination in Mammalian Cells.** *PLoS Genet* 2016, **12:**e1006410.
80. Ottaviani D, LeCain M, Sheer D: **The role of microhomology in genomic structural variation.** *Trends Genet* 2014, **30:**85-94.
81. Smith CE, Llorente B, Symington LS: **Template switching during break-induced replication.** *Nature* 2007, **447:**102-105.
82. Tsaponina O, Haber JE: **Frequent Interchromosomal Template Switches during Gene Conversion in S. cerevisiae.** *Mol Cell* 2014, **55:**615-625.

83. Mirkin EV, Mirkin SM: **Replication fork stalling at natural impediments.** *Microbiol Mol Biol Rev* 2007, **71:**13-35.

84. Moller HD, Parsons L, Jorgensen TS, Botstein D, Regenberg B: **Extrachromosomal circular DNA is common in yeast.** *Proc Natl Acad Sci U S A* 2015, **112:**E3114-3122.

85. Gresham D, Usaite R, Germann SM, Lisby M, Botstein D, Regenberg B: **Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus.** *Proc Natl Acad Sci U S A* 2010, **107:**18551-18556.

86. Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, Dutta A: **Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues.** *Science* 2012, **336:**82-86.

87. van Loon N, Miller D, Murnane JP: **Formation of extrachromosomal circular DNA in HeLa cells by nonhomologous recombination.** *Nucleic Acids Res* 1994, **22:**2447-2452.

88. Gaubatz JW: **Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells.** *Mutation Research/DNAging* 1990, **237:**271-292.

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

90. Rockman MV, Wray GA: **Abundant raw material for cis-regulatory evolution in humans.** *Mol Biol Evol* 2002, **19:**1991-2004.

91. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ: **Variable tandem repeats accelerate evolution of coding and regulatory sequences.** *Annual review of genetics* 2010, **44:**445-477.

92. Lee YCG, Karpen GH: **Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution.** *Elife* 2017, **6**.

93. Khost DE, Eickbush DG, Larracuente AM: **Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster.** *Genome Res* 2017.

94. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18:**188-196.

95. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al: **FlyBase 2.0: the next generation.** *Nucleic Acids Res* 2019, **47:**D759-D765.

96. Katoh K, Standley DM: **MAFFT: iterative refinement and additional methods.** *Methods Mol Biol* 2014, **1079:**131-146.

97. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

98. Yu G, Lam TT, Zhu H, Guan Y: **Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree.** *Mol Biol Evol* 2018, **35:**3041-3043.

99. McAllister BF, Werren JH: **Evolution of tandemly repeated sequences: What happens at the end of an array?** *Journal of molecular evolution* 1999, **48:**469-481.

100. Bailey TL, Johnson J, Grant CE, Noble WS: **The MEME Suite.** *Nucleic Acids Res* 2015, **43:**W39-49.

101. **BioCircos: Interactive Circular Visualization of Genomic Data using 'htmlwidgets' and 'BioCircos.js'. R package version 0.3.4.** [https://CRAN.R-project.org/package=BioCircos]

102. Shoura MJ, Gabdank I, Hansen L, Merker J, Gotlib J, Levene SD, Fire AZ: **Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in Caenorhabditis elegans and Homo sapiens.** *G3 (Bethesda)* 2017, **7:**3295-3303.

103. Usakin L, Abad J, Vagin VV, de Pablos B, Villasante A, Gvozdev VA: **Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in Drosophila melanogaster ovaries.** *Genetics* 2007, **176:**1343-1349.

104. Sproul JS, Maddison DR: **Sequencing historical specimens: successful preparation of small specimens with low amounts of degraded DNA.** *Mol Ecol Resour* 2017, **17:**1183-1201.

105. Chang CH, Larracuente AM: **Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome.** *Genetics* 2019, **211:**333-348.

106. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.

107. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biol* 2015, **16:**259.

FIGURE LEGENDS

Figure 1. Complex satellites in the heterochromatin of *D. melanogaster* and the simulans clade are in different locations. FISH image of mitotic chromosomes showing *Rsp-like* (red) and *1.688* (green) satellites. Chromosomes are counterstained using DAPI.

Figure 2: Euchromatic X-linked satellites are unevenly distributed across the X chromosome. (a.) A schematic illustrating terms frequently used in the text. We use ‘cytoband’ to reference large regions of the X chromosome that are defined by banding patterns in polytene chromosomes. We use ‘cluster’ to mean any distinct genomic locus containing the repeat of interest; typically, clusters contain several tandem repeats, although single-repeat clusters also exist. ‘Monomer’ refers to a single repeat unit; the example shown represents a *1.688* monomer. (b.**)** The x-axis shows position of *1.688* and *Rsp-like* satDNA clusters along the X chromosome. Counts shown on the y-axis indicate the number of repeat copies (*i.e.*, monomers) within a cluster. Each bar on the chart represents a cytological subdivision (*e.g.*, 1A, 1B, etc.) in which counts of all repeats are pooled.

Figure 3: Organization of cytoband 3F repeat cluster. Schematic of 3F cluster in *D. melanogaster* and the simulans clade, as well as the outgroup species *D. erecta*. Cluster is flanked by two genes, *echinus* and *roX1* (light green chevrons), with a TE insertion at the

distal side of the locus (purple chevrons). Complex satellite monomers are indicated by blue (*Rsp-like*) or orange (*1.688*) chevrons. Chevrons with dotted outline indicate sequences that were not annotated, but were determined manually by BLAST to be highly degenerated satellite monomers. Black dotted lines between species indicate shared repeats.

Figure 4. Comparison of phylogenetic patterns *1.688* and *Rsp-like* for *D. simulans*. Each terminal represents an individual repeat monomer from the X chromosome. Colored tip terminals indicate euchromatic repeats; gray tip terminals represent repeats from heterochromatic loci (defined as unassigned scaffolds in the assembly). Black rectangles indicate nodes with bootstrap support ≥ 90. Two regions in each tree are shown in greater detail to highlight differential phylogenetic patterns observed in euchromatic repeats of *1.688* and *Rsp-like*; arrows and dotted lines indicate relative position of enlarged regions in the tree. Branch lengths shown are proportional to divergence with both trees shown on the same relative scale. Sizes of the tips are scaled to reflect proportion of eccDNA reads mapping to a given variant, expressed as reads-per-million (RPM) (see eccDNA analysis). Maximum likelihood trees were inferred in RAxML with nodal support calculated following 100 bootstrap replicates.

Figure 5: All-species maximum likelihood trees of euchromatic *1.688* and *Rsp-like*. Each terminal represents an individual repeat monomer. All monomers from clusters with ≥three repeats were included in the analysis. Species identity is indicated by branch color. Major inter and intralocus expansions of satellites discussed in the text are labeled with gray arrows. For interlocus expansions in *Rsp-like*, the species involved are listed along with cytological bands that are represented by monomers within the expansion. The outgroup (*D. erecta*) is indicated by gray branches. Black rectangles indicate nodes with bootstrap support ≥ 90. Maximum likelihood tree was inferred in RAxML with nodal support calculated following 100 bootstrap replicates. Branch length is shown proportional to relative divergence with both trees on the same relative scale. See Figures S14–17 for added detail as to genomic location of terminals.

Figure 6. Junctions at new Rsp-like insertions in *D. simulans* and *D. mauritiana*. Junctions from a subset of the newer *Rsp-like* clusters (blue text/lines/boxes) are aligned and grouped into three types based on common signatures with nearby *1.688* monomers (orange text/lines/boxes). Type 1 is found in *D. simulans* while types 2 and 3 junctions are found in *D. mauritiana* (cytoband location of each cluster is indicated in the names at far left). Within each type, identical truncated *Rsp-like* monomers abut *1.688* at the same position in the *1.688* repeat monomer. In all three junction types, there is overlap between the two satellite sequences (black text) which, for at least the longer overlaps, potentially represents microhomology involved in the original insertion event. The second junction associated within and among these types is more variable ("var" in figure) with *Rsp-like* sequences abutting different positions of the *1.688* repeat or different unannotated sequences (gray boxes). The number of full length *Rsp-like* monomers as well as the lengths of truncated *Rsp-like* monomers, unannotated regions, and *1.688* sequences in this variable region are indicated for each cluster. Note that some clusters are nearly identical across this variable region (*e.g.*, Dsim7D and Dsim12F). The *1.688* sequences in the

region that would be sequential to those sequences at the conserved junctions (dark gray text above each junction type is the sequence within a specific *1.688* monomer) are indicated at the far right. Orange arrows in the first four *D. simulans* clusters indicate a duplication of the *1.688* sequences at the two junctions.

Figure 7. Proposed mechanisms of satDNA dynamics. Blue circles represent an ancestrally rare satellite (*i.e.*, *Rsp-like*), orange diamonds represent an abundant satellite present at many loci (*i.e.*, *1.688*), gray lines represent a fraction of a chromosome that spans many megabases. (a.) illustrates the microhomology-mediated birth of a hybrid repeat formed from the rare+common satellites, facilitating spreading of the rare satellite to loci where the abundant satellite is already present through processes illustrated by b–d. (b.) loci that are physically distant on a linear X chromosome may interact in three-dimensional space within the interphase nucleus, interlocus gene conversion of orange satellite repeats may then facilitate the spread of blue repeats. (c.) satellite DNAs are present on extrachromosomal circular DNAs, which may facilitate their spread to new loci. (d.) after new insertions of the blue satellite, entire mixed clusters may move as higher order units. The mechanisms illustrated in (b) and (c) could also be responsible for the generation of the hybrid repeat (a) and movement of higher order units (d). Not illustrated is the expansion or contraction of a repeat cluster at a given locus due to unequal exchange with a different cluster of the same repeat type.

Figure 8. Scatter plot of eccDNA RPM and genomic DNA RPM. Repeats in the genome are categorized into Other satellite (complex satellites except *1.688* and *Rsp-like*), LTR retrotransposon, non-LTR retrotransposon, DNA transposon and rolling-circle (RC) transposon and are shown in different colors. *Rsp-like* (shown in blue) and *1.688* (shown in orange) are indicated by arrows. Dotted lines represent the same abundance of eccDNA and genomic DNA such that dots above the dotted line indicate repeats that are enriched in eccDNA libraries relative to genomic controls.