TITLE : **Rapid turnover of centromeric DNA reveals signatures of genetic conflict in Drosophila**

AUTHORS :
Cécile Courret[1&], Lucas Hemmer[1], Xiaolu Wei[1], Prachi D. Patel[2], Bryce Santinello[2], Xuewen Geng[1], Ching-Ho Chang[3], Barbara Mellone[2,4], Amanda M. Larracuente[1&]

[&]**Corresponding authors:** ccourret@ur.rochester.edu, alarracu@bio.rochester.edu

AFFILIATION :

1. Department of Biology, University of Rochester; Rochester, New York, USA
2. Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, USA
3. Division of Basic Sciences, Fred Hutchinson Cancer Center, United States
4. Institute for Systems Genomics, University of Connecticut, Storrs, CT

ABSTRACT (150 words):

Despite their essential function in chromosome segregation, centromeres reside in rapidly evolving, repeat-rich genomic regions. Across organisms, centromeres are rich in selfish genetic elements like transposable elements and satellite DNAs that can bias their transmission through meiosis, but still need to contribute to centromere function. To gain insight into the balance between conflict and cooperation at centromeric DNA, we take advantage of the close evolutionary relationships within the *Drosophila simulans* clade – *D. simulans, D. sechellia,* and *D. mauritiana* – and their relative, *D. melanogaster*. We discovered dramatic centromere reorganization involving recurrent shifts between retroelements and satellite DNAs over short evolutionary timescales. We also reveal the recent origin (<240 Kya) of truly telocentric chromosomes in *D. sechellia,* where X and dot centromeres now sit on telomere-specific retroelements. This rapid centromere turnover is consistent with genetic conflicts in the female germline and has implications for centromeric DNA function and karyotype evolution.

INTRODUCTION

Cell division is an essential process for the viability of all organisms. Centromeres are chromosomal structures that are indispensable for faithful genome inheritance during cell division—they maintain sister chromatid cohesion and ensure proper chromosome segregation. Centromere defects can lead to the loss of genetic information and are associated with diseases (reviewed in [1]).

In eukaryotes, centromere identity is defined epigenetically by the presence of a centromere-specific histone H3 variant, CENP-A[2–4]. CENP-A, also known as CID in Drosophila, plays a central role in centromere function and identity and recruits kinetochore proteins, forming a macromolecular structure that allows spindle microtubule attachment[3]. Overexpression of CENP-A is associated with various cancers in humans[5]. In contrast to CENP-A protein, the role of the underlying DNA in centromere specification and function is not well-understood. In most species, centromeres are embedded in repetitive sequences[6], which makes it difficult to identify their precise organization. Despite the technical difficulties in studying such complex repetitive structures, recent studies highlight the importance of centromeric DNA in centromere stability and their impact on cell division and disease[7,8].

Centromeres vary widely in size and composition across species, from the point centromeres of *Saccharomyces cerevisiae* to the megabase-size arrays of the human centromeric α-satellite[6,9]. Although essential for proper chromosome segregation, both CENP-A and centromeric sequences are rapidly evolving, even among closely related species[10]. For example, among primate species, centromeric α-satellites belong to different repeat subfamilies[11,12]. One potential explanation of this paradox[10] is that genetic conflicts cause rapid centromere evolution[13]. Stronger centromeres can take advantage of the asymmetry in female meiosis to bias their transmission to the egg, rather than the polar body – a process known as centromere drive. Centromere drive has been observed in plants[14] and mammals[15–17]. Centromere strength is linked to its ability to recruit kinetochore proteins, which depends on centromeric DNA. For example, mouse centromeres with larger satellite DNA arrays recruit more centromeric proteins and thus increase their transmission through the female germline[18]. These satellite repeats thus may behave like "selfish" elements by promoting

centromeric chromatin expansion and causing segregation bias. If newly arisen satellite repeats can acquire more centromeric proteins, they will replace the old ones and could explain the rapid turnover of centromeric DNA sequences.

Satellite DNAs are not the only type of potentially selfish element occupying centromeres: transposable elements (TEs) are common features of centromeres in some fungi, plants, and animals[19]. TEs are generally selfish genetic elements, as they can proliferate within and spread between genomes, even when they come at a cost to their host[20]. While centromere function may not require any specific repeat sequence, some properties of satellite DNAs—*e.g.*, secondary structure[21,22], homogenized arrays, nucleosome-sized repeat units—may facilitate centromere maintenance and function[23]. TEs that insert in centromeres may interrupt otherwise homogenous arrays of satellites and affect centromere function[10,23]. However, the ubiquity of TEs at centromeres across a wide range of taxa suggest that they may instead play a conserved role in centromere specification, or even in centromere function (reviewed in[19,24]). Therefore, studying the evolutionary dynamics of centromeric DNA over short evolutionary timescales is important for understanding the balance between conflict and cooperation at centromeric DNA.

The small, but complex genomes of *Drosophila* species make them excellent models for the study of centromere function and evolution. In *Drosophila melanogaster,* centromeres correspond to islands of complex DNA highly enriched in retroelements and flanked by simple tandem satellite repeats[25]. While each centromere has a unique organization, they all share only one common component: a non-LTR retroelement called *G2/Jockey3*. *G2/Jockey-3* is also present in the centromeres of a closely related species, *D. simulans,* suggesting that it could be a conserved feature of *Drosophila* centromeres. While recent reports suggest that *D. melanogaster* and *D. simulans* centromeric regions have distinct satellite repeats[26], we do not know the precise organization of centromeres outside of *D. melanogaster.*

Here we combine (epi)genomic and cytologenetic approaches to study the evolutionary dynamics of centromeres in closely related species of the *simulans* clade - *D. simulans, D. sechellia* and *D. mauritiana*. These species diverged from each other only 240,000 years ago, and from *D. melanogaster* 2.4 million years ago[27,28], allowing us to study centromere evolution on two different

timescales at high resolution. We discover that none of the *D. melanogaster* retroelement-rich centromeres are conserved in the *D. simulans* clade. Instead, two complex satellites – a *365-bp* and a *500-bp* tandem satellite repeat – invaded the centromeres of these species. The centromere-associated *G2/Jockey-3* retroelement remains active in one of the lineages (*D. simulans*) but not the others. We also discover the origination of telocentric chromosomes in *D. sechellia,* where the centromeres of chromosomes *X* and *4* now sit on retroelements with telomere-specific functions. These replacements imply that centromeres can shift their composition rapidly, and between categorically different sequence types: TEs and satellite DNAs. The only chromosomes that do not show these categorical shifts in composition are the Y chromosomes, which have male-specific transmission. This suggests that the selection forces driving rapid centromere evolution are female-specific, consistent with recurrent genetic conflicts over transmission through the female germline. Our comparative study of detailed centromere organization has implications for the roles of retroelements and satellites in centromere function and evolution, and karyotype evolution.

RESULTS

Satellite invasion of simulans clade centromeres

To identify the detailed organization of centromeres in the *D. simulans* clade we performed CUT&Tag[29] on embryos from each species (*D. simulans, D. sechellia,* and *D. mauritiana*) using a CENP-A antibody. The resulting reads were mapped to versions of each species' genome assembly with improved representation of heterochromatic regions[30]. Because centromeres sit in highly repetitive genome regions, we analyzed uniquely- and multi-mapping reads independently (Figure 1, SupFig1-3). We identified centromere candidates as the top CENP-A-enriched contigs based on the peak density whose calls were consistent between replicates (irreproducible discovery rate [IDR] < 0.05, SupTable1 and SupFig4). We also used an assembly-free analysis to detect the enrichment of complex repeats in the CENP-A CUT&Tag reads (see Methods). We validated our approach by performing the CUT&Tag on a *D. melanogaster* strain, which recovered the same centromere islands as in Chang, Palladino, and Chavan et al. [25] (SupFig5).

Like *D. melanogaster*, all three simulans clade species have a pair of large metacentric "major" autosomes (chromosomes *2* and *3*), a pair of small autosomes (chromosome *4;* referred to as the "dot" chromosome), and a pair of sex chromosomes (*X* and *Y*). For each species, there were five contigs that were consistently among the most CENP-A-enriched contigs (SupFig4). We considered these contigs to be the centromere candidates for each chromosome (SupTable2). None of the *simulans* clade centromere candidates we identified were like *D. melanogaster* centromeres, suggesting a turnover in centromere identity in the 2.4 My since these species diverged. Instead, both our assembly-based (Figure 1A,D,G, SupFig1-3) and assembly-free (Figure 1B,E,H) approaches identify the *500-bp* complex satellite among the most CENP-A enriched sequences (Figure 1). The centromere candidate contigs for the major autosomes in *D. simulans*, *D. mauritiana* and *D. sechellia* (Figure 1A,D,G, SupFig1-3) and the X chromosome in *D. simulans* and *D. mauritiana* (Figure 2, SupFig1-3) are mainly comprised of the *500-bp* satellite repeat. This complex satellite was previously identified as being associated with the centromeres in *D. simulans*[26].

The *500-bp* is enriched in, but not specific to, *simulans* clade centromeres, as we also find it in the proximal pericentromeric regions. The CENP-A domain appears restricted to a subset of the *500-bp* satellite array spanning 50 kb to 200 kb (Figure 1A, D, G and 2). This is similar to human centromeres, where the CENP-A domain sits on a subset of α-satellite repeats within an array[31]. We also identified a second complex satellite associated with centromere candidates, which we named the *136-bp* satellite. While less abundant, *136-bp* is interleaved with the *500-bp* satellite and is associated with the same centromeres (Figure 1 and 2, SupFig6A).

We validated our observations using a cytogenetic approach with IF-FISH on mitotic chromosome spreads from larval brains using Oligopaints targeting each complex satellite[32]. We confirmed the localization of centromeric protein (CENP-C) on the *500-bp* (Figure 1C, F, and I) and *136-bp* (SupFig6A) satellites. While the major autosomal centromeres are mainly composed of the same complex satellites in the three species, the distal pericentromere appears more divergent. In *D. simulans* and *D. mauritiana,* the major autosomal pericentromeres contain the *dodeca* satellite (Figure 1C,F), while in *D. sechellia* they contain the *Rsp-like* satellite (Figure 1I). We also found the *Rsp-like* satellite on the X pericentromere of *D. simulans* (Figure 2A), although it was absent in *D. mauritiana* (Figure 2B)[33,34]. The combination of satellites flanking the CENP-A domain (Figure 1C, F, I and Figure 2) allows us to assign the *500-bp* rich contigs to either the major autosomes (Figure 1A, D and G) or the X chromosome (Figure 2). Unfortunately, we cannot morphologically distinguish between the chromosomes *2* and *3* because of their similarity.

We used a BLAST approach to explore origins of the *500-bp* and *136-bp* centromeric complex satellites and did not find any evidence of their presence outside of the *D. simulans* clade, even as single copy sequences (SupTable3, SupTable4). For example, in *D. melanogaster*, the best hit had 85% identity with the *500-bp* consensus sequence but only covered 106 bp of the query (SupTable3). This suggests that these satellites emerged after the divergence between *D. melanogaster* and the *D. simulans* clade 2.4 Mya[27,28], although it is possible that the primary sequence emerged earlier but was lost in *D. melanogaster* and spread in the *D. simulans* clade species. In either case, these satellites recently invaded the *D. simulans* clade centromeres (SupFig7).

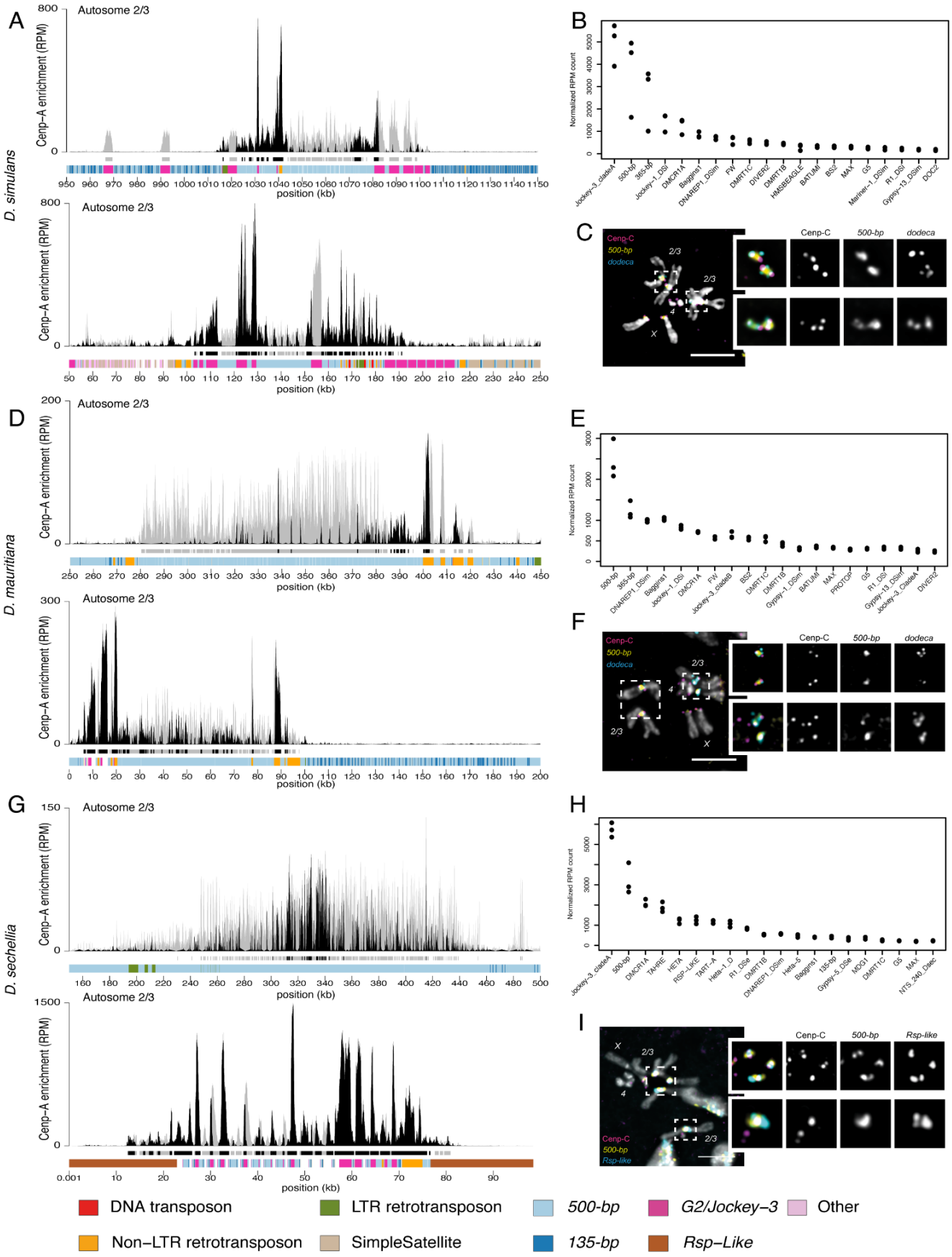

A
B
C
D
E
F
G
H
I
D. simulans
D. mauritiana
D. sechellia
Autosome 2/3
Cenp−A enrichment (RPM)
position (kb)
Normalized RPM count
Cenp-C
500-bp
dodeca
Rsp-like
DNA transposon
LTR retrotransposon
500-bp
G2/Jockey−3
Other
Non−LTR retrotransposon
SimpleSatellite
135-bp
Rsp−Like

Figure 1. Centromeres of chromosomes *2* and *3* in *D. simulans*, *D. sechellia,* and *D. mauritiana* are predominantly *500-bp* satellite. (A,D,G) CENP-A CUT&Tag enrichment on the centromere candidates for the major autosomes (*2* and *3*) of *D. simulans* (A), *D. mauritiana* (D) and *D. sechellia* (G). The y-axis represents normalized CENP-A enrichment in Reads Per Million (RPM). Black and gray plotted lines represent the enrichment based on uniquely and multi mapping reads, respectively. The black and gray tracks below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads, respectively. The colored cytoband track at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure. (B,E,H) Assembly-free analysis showing the normalized enrichment score (in RPM) of CENP-A for complex repeats, including transposable elements and complex satellites across all centromeres. The Top 20 most enriched repeats are represented for *D. simulans* (B), *D. mauritiana* (E) and *D. sechellia* (H). (C,F,I) IF-FISH on mitotic chromosomes from larval brains with CENP-C antibody and *500-bp* and *dodeca* probes, for *D. simulans* (C) and *D. mauritiana* (F) or *500-bp* and *Rsp-like* probes for *D. sechellia (I)*. The insets represent a zoom on each major autosome centromere. Bars represent 5 µm.

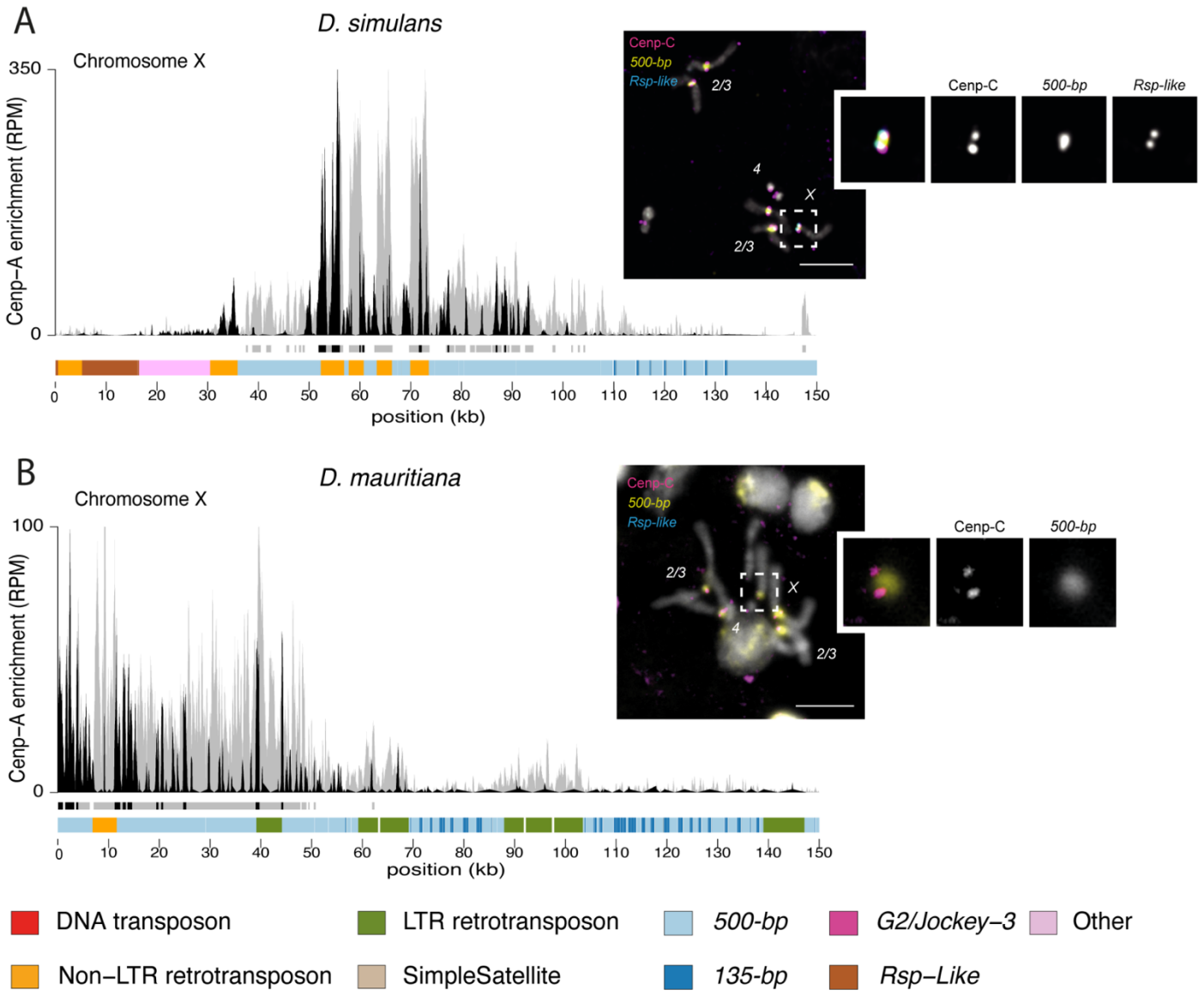


Figure 2. X chromosome centromeres in *D. simulans* and *D. mauritiana* are enriched in *500-bp* satellite. The left panel shows the CENP-A CUT&Tag enrichment on the X centromere candidate in *D. simulans* (A) and *D. mauritiana* (B). The y-axis represents the normalized CENP-A enrichment in RPM. Black and gray plotted lines represent the enrichment based on uniquely and multi mapping reads, respectively. The black and gray tracks below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads, respectively. The colored cytoband at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure. The right panel shows IF-FISH on mitotic chromosomes from larval brains with CENP-C antibody and *500-bp* and *Rsp-like* probes. The inset represents a zoom on each X chromosome centromere. Bars represent 5 µm.

Dot chromosome centromeres are enriched with a chromosome-specific complex satellite

In *D. simulans* and *D. mauritiana,* the centromere of the small autosomal dot chromosome (*i.e.,* Chromosome *4*) contains a different complex satellite: the *365-bp* satellite (Figure 3). This repeat is consistently enriched in CENP-A chromatin in both our assembly-based (Figure 3) and assembly-free (Figure 1B and E) approaches. The CENP-A domain is restricted to the *365-bp* satellite and flanked by the *AATAT* satellite on at least one side (Figure 3), which is confirmed by our FISH with CENP-C IF on chromosome spreads (Figure 3 insets). Unlike the *500-bp* satellite, *365-bp* is specific to the dot chromosome centromere. We do not find evidence of the *365-bp* satellite outside of one CENP-A enriched contig in each assembly (Figure 3), consistent with the FISH signals (Figure 3 insets).

We used a BLAST-based approach to explore the origin of the *365-bp* satellite and did not find evidence of this satellite outside of the *D. simulans* clade species (SupTable5). For example, in *D. melanogaster,* the best hit had 82% identity with the *365-bp* consensus sequence but was only 57 bp long (SupTable5) suggesting that, like the *500-bp* satellite, the *365-bp* satellite emerged after the split with *D. melanogaster* and likely invaded dot centromeres in the ancestor of the *D. simulans* clade (SupFig7).

Interestingly, *365-bp* was lost from *D. sechellia:* we did not find evidence of this satellite, even as a single copy sequence, neither within the *D. sechellia* genome assembly nor the Illumina raw reads from genomic DNA (SupTable5). Based on our assembly-free approach, the *365-bp* satellite does not appear enriched among the CENP-A reads in *D. sechellia* (Figure 1H). We also did not obtain any signal from the FISH on mitotic chromosomes using Oligopaints targeting the *365-bp* satellite in this species, while the pericentromeric *AATAT* satellite appears to be conserved (SupFig6B).

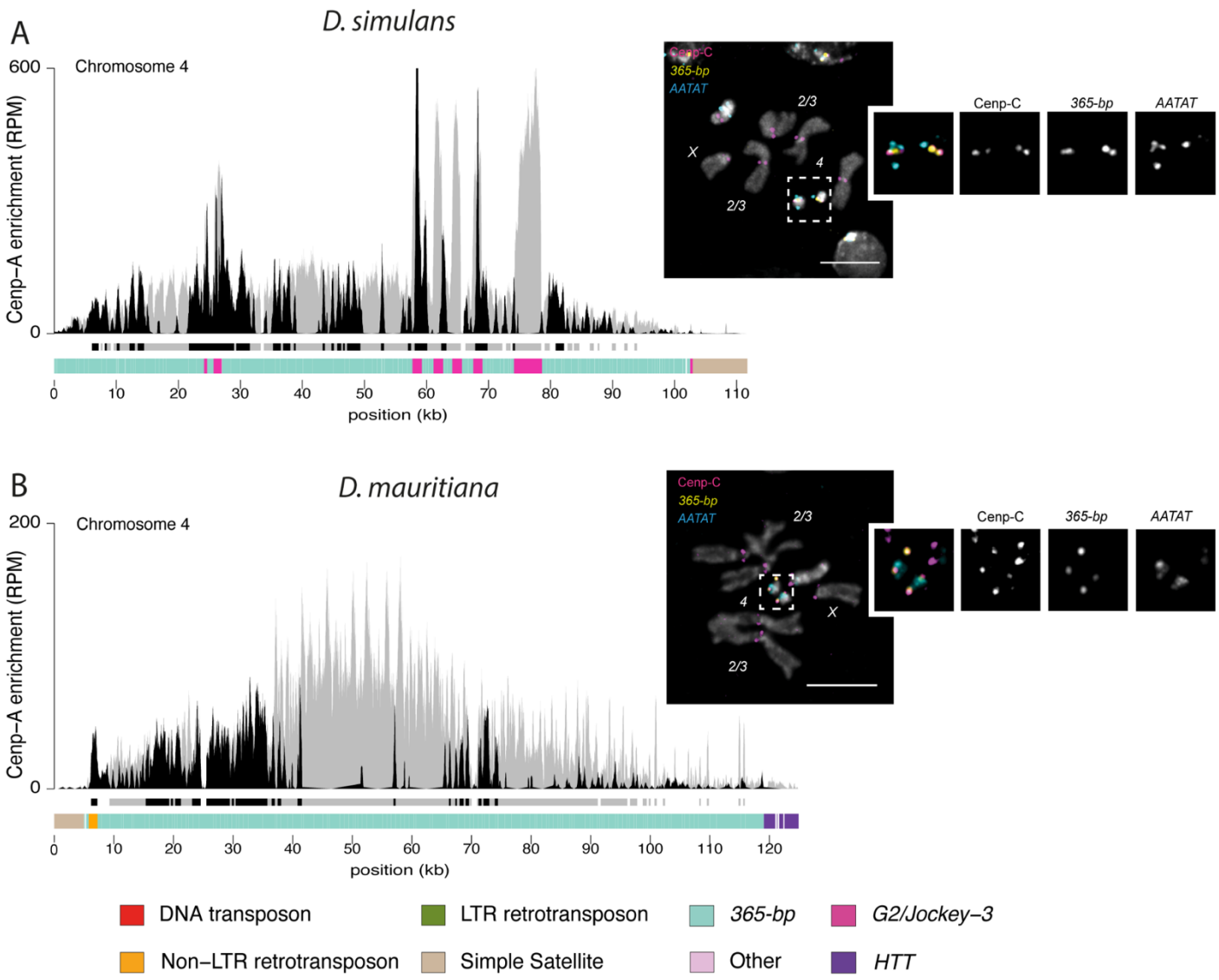


Figure 3. Dot chromosome centromeres in *D. simulans* and *D. mauritiana* are enriched in *365-bp* satellite. The left panel represents the CENP-A CUT&Tag enrichment in *D. simulans* (A) and *D. mauritiana* (B). The y-axis represents the normalized CENP-A enrichment in RPM. Black and gray plotted lines represent the enrichment based on uniquely and multi-mapping reads, respectively. The black and gray tracks below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi-mapping reads, respectively. The colored cytoband track at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure. The right panel represents the IF-FISH on mitotic chromosomes from the larval brain with CENP-C antibody and *365-bp* and *AATAT* probes. The insets represent a zoom on each dot chromosome centromere. Bars represent 5 µm.

Centromere shifts to telomere-specialized retroelements: truly telocentric chromosomes in D. sechellia

In *D. sechellia,* the dot and X chromosome are different from those of *D. simulans* and *D. mauritiana*. We did not identify any *500-bp*-enriched contig that might correspond to the X chromosome centromere and *365-bp* is missing from the *D. sechellia* genome.

Instead, we identified two *D. sechellia* contigs that are significantly enriched for CENP-A containing an array of non-LTR retroelements well known for their role at telomeres: *Het-A*, *TART* and *TARHE* (also known as the *HTT* elements)[35]. The *HTT* elements are also among the most CENP-A-enriched elements in our assembly-free approach (Figure 1H). *Drosophila* species lack telomerases; instead, telomere size and integrity is maintained by the transposition activity of *HTT* retroelements. *HTT* elements have specialized functions at telomeres of some *Drosophila* species, including *D. melanogaster* and the *D. simulans* clade[35].

On one *HTT*-CENP-A enriched contig, the *HTT* domain is flanked by the *500-bp* satellite, suggesting that it corresponds to the X chromosome centromere (Figure 4C). However, CENP-A is enriched on the *HTT* domain instead of the flanking *500-bp* satellite (Figure 4A), suggesting a shift in the centromere to the retroelements that normally occupy the telomere. Similarly, on the second *HTT*-CENP-A enriched contig, a simple *ATAG* satellite flanks the CENP-A domain (Figure 4B). This satellite is associated with both the X and dot chromosome[36], and thus we deduced that this second contig corresponds to the dot chromosome centromere.

To validate our observations, we designed Oligopaints targeting the *HTT* array on the X and dot chromosome centromere candidates in *D. sechellia.* The IF-FISH on mitotic chromosomes from larval brains confirmed that the centromeric protein CENP-C is indeed associated with the *HTT* domain on both the X and dot chromosomes. The *500-bp* satellite appears adjacent to the *HTT* on the X chromosome (Figure 4C).

To visualize these regions at higher resolution, we performed IF-FISH on stretched chromatin fibers using a CENP-A antibody and Oligopaints targeting the *500-bp* satellite and the *HTT* elements. These fibers confirm that CENP-A nucleosomes are seated on the *HTT* domain and are

flanked by, but do not overlap, the *500-bp* satellite (Figure 4D). The chromatin fibers appear to end just after the CENP-A/*HTT* signal, strongly suggesting that the centromere is on a telomeric *HTT* array, making these chromosomes truly telocentric (Figure 4D). We also observed patterns from stretched chromatin fibers consistent with our predictions for the other chromosome centromeres (SupFig8).

Interestingly, the dot chromosome centromere of *D. mauritiana* is flanked by the *AATAT* satellite on one side and by the *HTT* on the other side (Figure 3B). Unfortunately, the contig is not long enough to establish how long the *HTT* domain is after the centromere, but it suggests that in *D. mauritiana,* both centromeric and telomeric domains are very close to each other.

It was very surprising to find the centromeric protein associated with telomeric sequences, as centromeres and telomeres are generally domains with distinct functions. Although both the X and the dot chromosomes were considered to be acrocentric chromosomes based on the similarity in karyotype with *D. melanogaster*[37,38], our high-resolution approach allowed us to reveal truly telocentric chromosomes for the first time in these species.

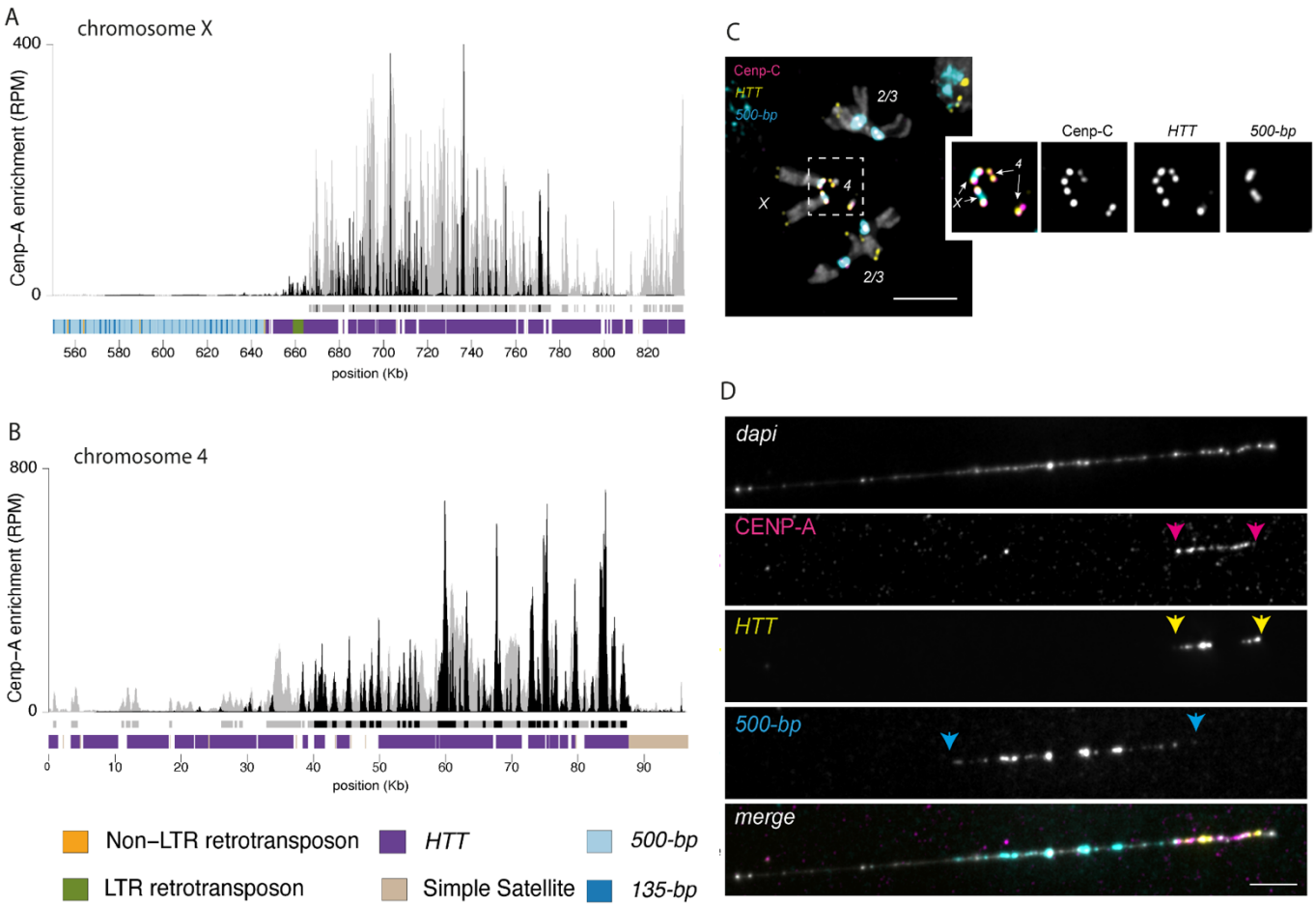


Figure 4. The Dot and X chromosome centromere in *D. sechellia are telocentric*. CENP-A CUT&Tag enrichment along the X **(A)** and dot **(B)** chromosome centromeres. The y-axis represents the normalized CENP-A enrichment in RPM. Black and gray plotted lines represent the enrichment based on uniquely and multi-mapping reads, respectively. The black and gray tracks below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi-mapping reads, respectively. The colored cytoband track at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure. **C)** IF-FISH on mitotic chromosomes from the larval brain with CENP-C antibody and *500-bp* and *HTT* probes. The inset represents a zoom on the X and dot chromosome centromeres. Bar represents 5 um. **D)** IF-FISH on chromatin fibers from the larval brain with CENP-A antibody and *500-bp* and *HTT* probes, representing the telocentric X chromosome of *D. sechellia*. Bar represents 5 µm.

The Y chromosome centromeres are unusual.

The Y chromosome centromeres are unique in their composition and organization compared to the rest of the genome. Unlike the other chromosomes, we did not identify any complex satellites associated with the Y chromosome centromere. Instead, CENP-A is enriched in a region with high density of transposable elements (Figure 5). While each species' Y chromosome centromeres are mainly enriched in retroelements, they each have a unique composition (Figure 5, SupTable6). For example, the most abundant elements associated with the Y centromere are *HMSBEAGLE* and *Jockey-1* in *D. simulans*, *mdg4* in *D. mauritiana* and *R1* and *G2/Jockey3* in *D. sechellia* (SupTable6). Interestingly, both the *D. simulans* and *D. sechellia* centromeric sequences form higher order repeats, but not *D. mauritiana* (SupFig9).

To validate our candidate Y centromeres, we designed Oligopaints specific to the Y contig of each species (*cenY*). We performed IF-FISH on mitotic chromosomes with a CENP-C antibody and the Oligopaint targeting the putative Y centromeres. Our Oligopaints give a signal specific to a unique region of the Y chromosome which consistently co-localizes with the CENP-C signal (Figure 5), confirming the Y chromosome centromeres.

While simple satellites are present within the pericentromeric region of all the other chromosomes, we do not find any simple satellites in the flanking region of the Y centromere (Figure 5). This is surprising, especially given that these Y chromosomes are highly enriched in simple satellites in general[39,40].

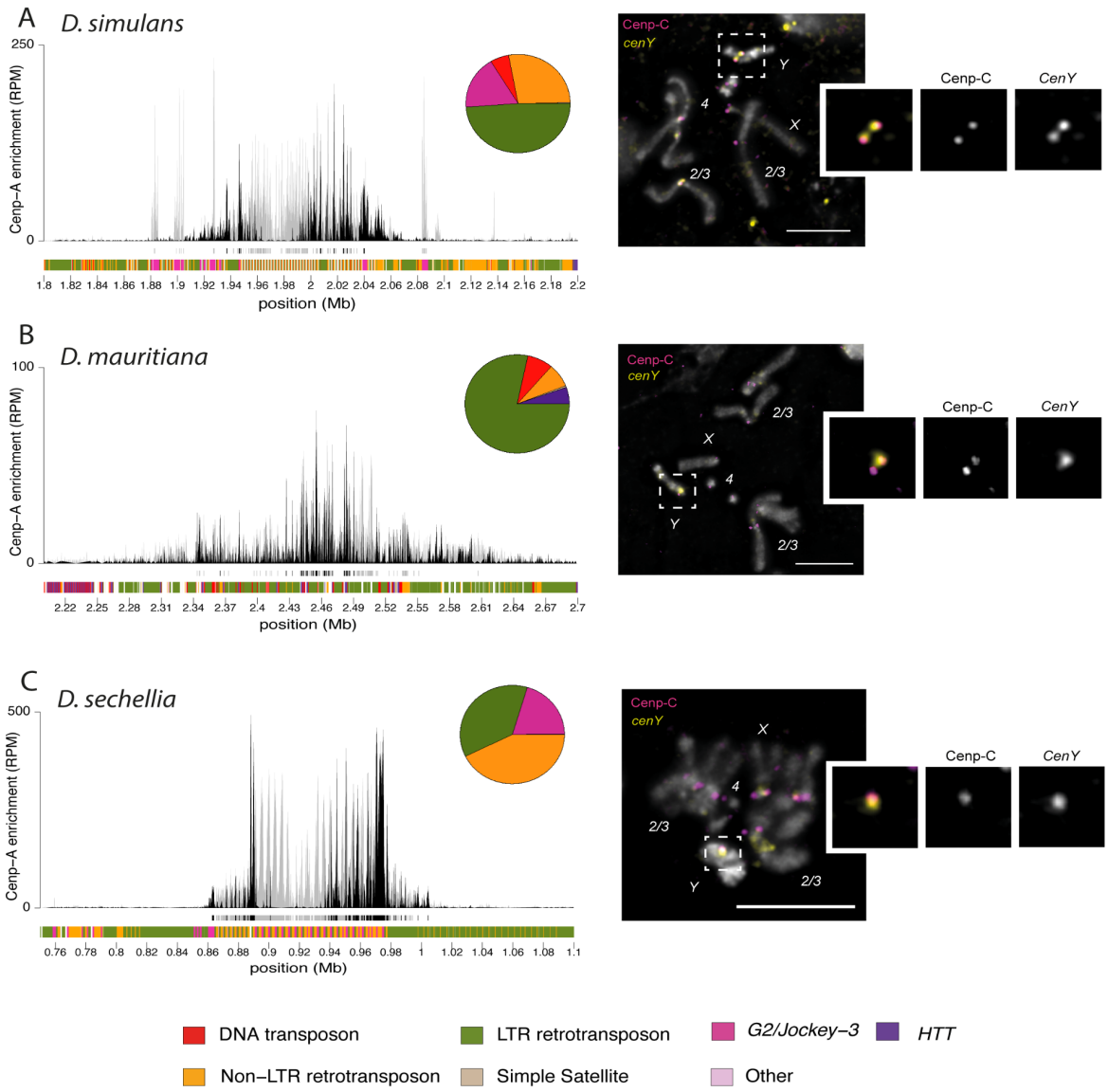


Figure 5 : **The Y chromosome centromeres of *D. simulans*, *D. mauritiana* and *D. sechellia* are rich in transposable elements**. The left panel shows the CENP-A CUT&Tag enrichment for the Y centromere of *D. simulans* (A), *D.mauritiana* (B) and *D. sechellia* (C). The y-axis represents the normalized CENP-A enrichment in RPM. Black and gray plotted lines represent the enrichment based on uniquely and multi-mapping reads, respectively. The black and gray tracks below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi-mapping reads, respectively. The colored cytoband track at the bottom of the plot shows the repeat organization. The pie chart on the top represents the repeat composition of the CENP-A domain. The color code of the cytoband and pie chart is shown in the legend at the bottom of the figure. The right panel shows the IF-FISH on mitotic chromosomes from the larval brain with CENP-C antibody and *cenY* Oligopaints specific to each species' centromere. The insets represent a zoom on each Y chromosome centromere. Bar represents 5 µm.

G2/Jockey-3 is associated with centromeres within the *D. simulans* clade

In *D. melanogaster,* the only common sequence among all centromeres is *G2/Jockey-3*[25]. We asked if this element was also found within the *simulans* clade centromeres. In *D. simulans*, *G2/Jockey-3* is the most enriched repeat among the CENP-A reads (Figure 1B). We identified *G2/Jockey-3* insertions in each centromere except for the X chromosome (Figure 2A), where it directly flanks the centromere (Figure 6C). We confirmed the presence of *G2/Jockey-3* at each centromere by IF-FISH on mitotic chromosomes (Figure 6C). In *D. sechellia*, *G2/Jockey-3* is also the most enriched repeat in CENP-A chromatin (Figure 1H); however, we only detect it on the Y chromosome and one of the autosomal centromeres (Figure 1G, 5C, 6C). Similarly, in *D. mauritiana, G2/Jockey-3* is associated with only one of the autosomal centromeres (Figure 1D, 6C), and is less enriched than in the two other species (Figure 1E). This suggests that the association of *G2/Jockey-3* with the centromere was lost.

To better understand the evolutionary history of this specific retroelement, we inferred the phylogeny for all *G2/Jockey-3* ORFs in the *D. melanogaster* clade assemblies. *G2/Jockey-3* has two open reading frames (ORFs), but we only used ORF2 for inferring phylogenies, as ORF1 is more evolutionarily labile across species[41]. While all *D. melanogaster G2/Jockey-3* insertions cluster together in a unique clade, the *D. simulans* clade insertions form two different clades, which we designate as clade 'A' —with sequences more closely related to *D. melanogaster G2/Jockey-3* — and clade 'B' (Figure 6A, SupFig10). Within each clade, insertions largely form species-specific clusters. All centromeric insertions are part of the clade 'A'. Only the *D. simulans* clade 'A' consensus has a conserved ORF. Like *D. melanogaster, G2/Jockey-3* clade 'A' insertions are enriched in centromeres (Figure 6B). That is, 53% of all *G2/Jockey-3* insertions are centromeric in *D. simulans* and *D. sechellia,* which is more than expected if these TEs were randomly distributed in the genome (Fisher's exact tests: $P_{sim} < 10^{-16}$; $P_{sec} < 10^{-16}$;). The enrichment is less pronounced in *D. mauritiana* (17%; $P_{mau} = 0.0567$). However, the consensus ORF is incomplete in *D. sechellia* and *D. mauritiana*, implying that most clade 'A' *G2/Jockey-3* copies are inactive in these species, in line with their inconsistent association with centromeres. These findings suggest that a subset of *G2/Jockey-3* elements likely had centromere-biased insertion activity in the *D. melanogaster* clade ancestor. This activity may have continued after the speciation event between *D. melanogaster* and the *D. simulans* clade but was lost in *D. sechellia* and *D. mauritiana* lineages, resulting in the

inability of *G2/Jockey-3* to jump into centromeres. While the clade 'B' appears to have been recently active in the simulans clade, none of the insertions are centromeric. This clade was either lost from *D. melanogaster* or may have been introduced into the *D. simulans* ancestor from a horizontal transfer event. The latter appears to be more likely as we find fragmented copies of *G2/Jockey3* from *D. yakuba* that cluster with clade 'B', although we do not have sufficient node support to draw strong conclusions about the origins of this clade. Either way, our data suggest that the clade 'A' variant of *G2/Jockey-3* targeted the centromeres for insertion despite the complete turnover in centromere composition between *D. melanogaster* and the *D. simulans* clade species, suggesting that this element may preferentially target centromeric chromatin rather than particular DNA sequences.

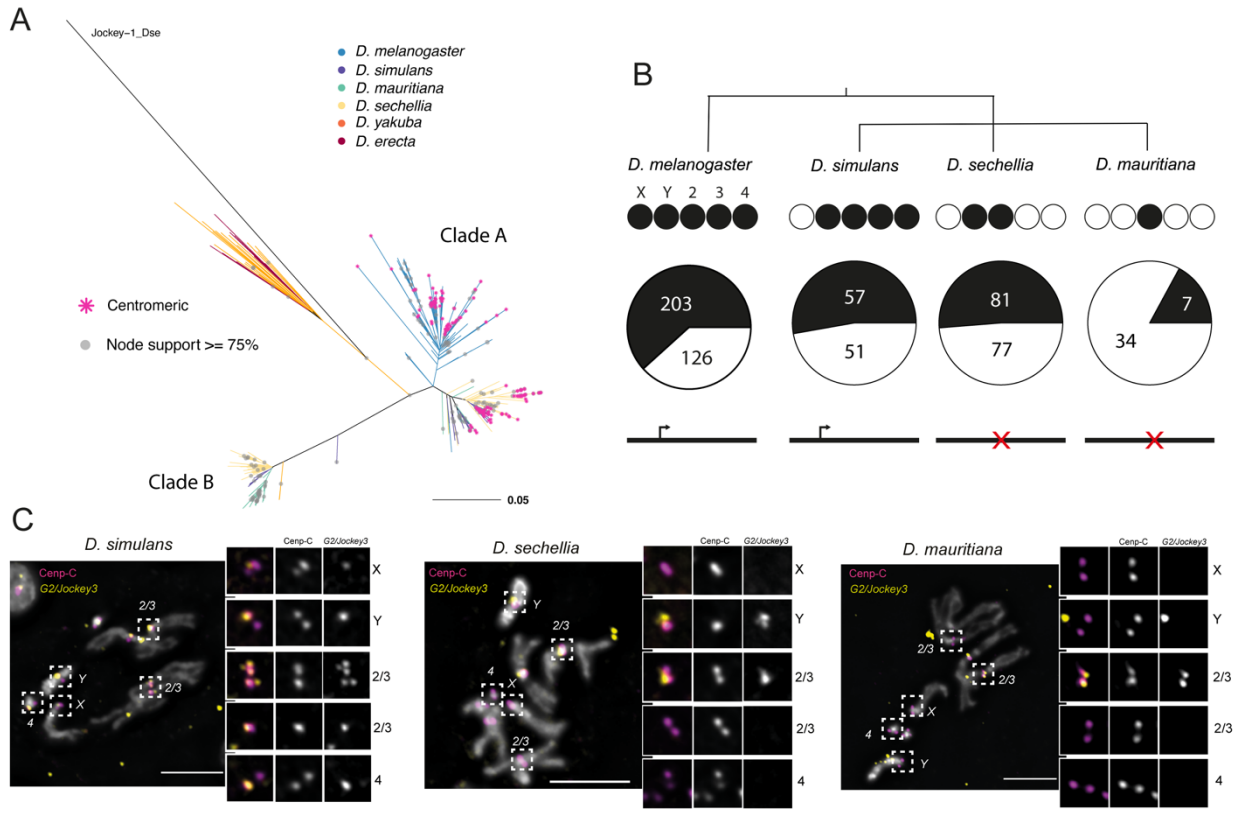


Figure 6: ***G2/Jockey*-3 is associated with the centromeres within the *D. simulans* clade. A)** Maximum likelihood phylogenetic tree of *G2/Jockey-3* ORF2 from *D. melanogaster, D. simulans, D. sechellia, D. mauritiana, D. yakuba,* and *D. erecta. G2/Jockey-3* within the *simulans* clade species diverged into two different clades, one that is more closely related to the *D. melanogaster* elements (clade 'A') and one that is more divergent (clade 'B'). Centromeric insertions are indicated by a pink * at the tip of the branch. We do not know centromere identity in *D. yakuba* and *D. erecta*. **B)** ORF2 conservation analyses of the clade 'A' *G2/Jockey-3* centromere-associated clade. The circles below the species name represents each centromere. Centromeres containing *G2/Jockey3* insertions (based on CUT&Tag and FISH) are shown in black. The pie chart represents the proportion of centromeric (black) and non-centromeric (white) insertions among the clade 'A' *G2/Jockey3* within each species' genome, where we indicate the number of insertions within the pie charts. The consensus sequence of *G2/Jockey3* ORFs is schematized below the pie chart, indicating that only *D. melanogaster* and *D. simulans* consensus sequences have an intact ORF2. **C)** IF-FISH on mitotic chromosomes from the larval brain with CENP-C antibody and *G2/Jockey3* probes showing consistent centromere-association in *D. simulans*, but not in *D. mauritiana* and *D. sechellia*. In *D. simulans*, the *G2/Jockey-3* insertions on the X chromosome are adjacent to the CENP-A domain, rather than within. The inset represents a zoom on each centromere. Bars represent 5µm.

DISCUSSION

Our combination of genomic, epigenetic, and cytological approaches identified the precise composition and organization of each centromere in the *simulans* clade species and revealed a rapid reorganization of centromeres to categorically different repeats over short evolutionary timescales. While *D. melanogaster* centromeres primarily correspond to retroelement-rich islands flanked by simple satellites, in the *simulans* clade centromeres are composed by two complex satellites: *500-bp* and *365-bp*. We do not find evidence of these satellites outside of the *D. simulans* clade, suggesting that they recently expanded and invaded the *D. simulans* clade centromeres. These events resulted in a complete turnover of centromere composition — from retroelements to complex satellite-based centromeres — in less than 2.4 million years.

In the last decade, several studies have shed light on the rapid evolution of centromere sequences in a wide range of species[9]. However in most studies, centromeres generally remain associated with the same type of repeat element, shifting to different variants of related repeats (either retroelements or satellites)[43–50] thus maintaining a certain homogeneity in the type of centromeric sequence among closely related species. For example, human and its closely related species — chimpanzee, orangutan, and macaque — diverged about 7 million years ago but all centromeres are populated by different subfamilies of the α-satellite repeat[44,45]. *Arabidopsis* species, *A. thaliana* and *A. lyrate,* also experienced a turnover of centromere sequences in the 5-6 My since their divergence[51], but between related satellites[46]. In our study of *Drosophila*, we highlight a complete turnover of centromere sequences to categorically different repeats – satellites and retroelements – between closely related species.

We also reveal a more dramatic shift in centromere composition in *D. sechellia*. The X and dot chromosomes of the *melanogaster* species are classified as acrocentric based on cytological observations of mitotic chromosomes (reviewed in[37]). Here, we use epigenetic profiling and high resolution cytology to reevaluate this assumption, allowing us to distinguish between chromosomes with independent, but nearby centromere and telomere domains (like in mouse where centromeres are positioned 10–1,000 kb away from the telomere[52,53]), and truly telocentric chromosomes where both occupy the same repetitive array. In *D. sechellia*, the CENP-A domain occupies the array of

telomeric *HTT* elements, which are known for their function in maintaining telomere length in *Drosophila* species. These centromere replacement events are specific to *D. sechellia* and thus likely occurred in the past 240,000 years. While the *500-bp* satellite remains adjacent to the telocentromeric domain on the X chromosome (Figure 4A-C), the *365-bp* satellite was lost from the *D. Sechellia* dot chromosome. We cannot infer if the centromere shift to the *HTT* was a consequence or cause of the loss of the *365-bp* satellite. However, the X chromosome observation makes the former scenario more likely. While in *D. sechellia* X and dot chromosomes are truly telocentric (Figure 4D), we think that in *D. simulans* and *D. mauritiana* the X and dot chromosomes are nearly telocentric, with CENP-A occupying different repeats very close to the telomeres. Indeed, the *D. mauritiana* dot centromere sits on the *365-bp* satellite, but the *HTT* elements are adjacent (Figure 3B).

Our observations raise important questions regarding the respective roles of centromeres and telomeres in chromosome biology as well as their functional association. Interestingly, in fission yeast the telomere bouquet is essential for spindle formation through telomere-centrosome contacts. However, if the telomere bouquet is disrupted, centromere-centrosome contacts can rescue the spindle defect, suggesting that centromeres and telomeres have functional similarities and interchangeable roles[54]. Similarly in mice, one of the shelterin complex proteins that is essential for telomere function (TRF1) is also required for centromere and kinetochore assembly[55]. In the case of *D. sechellia, HTT* elements with historical telomere-specific functions, now need to also carry out centromere functions, at least at the structural level.

In *D. melanogaster,* all centromeres are composed of retroelements, but each centromere has a unique sequence composition. The only common element between centromeres is *G2/Jockey3*. While it is tempting to speculate that the centromere occupation of *G2/Jockey-3* in *D. melanogaster* and *D. simulans* suggests that it plays a functional role at centromeres, we propose that *G2/Jockey-3* is instead an opportunistic selfish genetic element. Centromeres provide 'safe havens' for TEs to insert far from protein-coding genes and with little opportunity for deleterious ectopic recombination[24,56,57]. Centromeres may tolerate TEs that contribute positively to a proper chromatin and transcription environment for centromere function, as RNA transcripts are important for centromere assembly across species [58–62]. TEs may not only limit harm to their host if they

insert in centromeres, but they may also escape host defenses if they insert in CENP-A nucleosomes[63]. However, TEs that invade centromeres may be generally deleterious, as a high density of active TEs may inactivate centromeres through heterochromatinization[23,64]. *G2/Jockey-3* seems to have lost its activity in *D. mauritiana* and *D. sechellia* but appears to be recently active in *D. simulans.* Because of categorical differences between species in underlying centromeric DNA, we believe that the consistent association between *G2/Jockey-3* and centromeres in *D. melanogaster* and *D. simulans* represents independent insertion events rather than conservation of individual elements. *G2/Jockey-3* seems to target centromeres regardless of their underlying sequence composition, suggesting that this element may target centromeric chromatin itself rather than a specific sequence. Retroelements may specifically target centromeres in other species such as maize [65,66] and *Arabidopsis* [45,66,67]. Transformation experiments in Arabidopsis showed that the centromere-associated *Tal1* retroelement from *A. lyrata* is able to target *A. thaliana* centromeres[67] despite divergent (30%) centromeric satellites in these species[68].

Klein and O'Neill proposed that retroelements could play an important role in fueling rapid centromere evolution[24]. Retroelement transcription can favor the recruitment of new insertions at neocentromeres, recruiting more CENP-A to stabilize the centromere. Recurrent insertions may also facilitate the emergence, or the spread, of satellites, which if favored by selection or selfish dynamics, can become the major component of centromeres. Centromeres may then cycle between retroelement-rich and satellite-rich domains through repeated bouts of retroelement invasion, followed by satellite birth, and satellite invasion events (Figure 7B).

In this study, we describe two types of centromere turnover events (Figure 7A): *1)* the invasion of complex satellites in the ancestor of the *D. simulans* clade (2.4 Mya), replacing the previously retroelement-rich and satellite-poor *D. melanogaster* centromeres; and *2)* and a shift to truly telocentric chromosomes, with satellite-free centromeres, comprised only of telomere-specialized retroelements (in <240 Kya). Such dramatic shifts in centromere composition raise questions about the role of DNA sequences in centromere function and the dynamic processes driving such shifts.

The centromere drive hypothesis predicts an evolutionary arms race between centromere sequences and centromeric proteins, and might explain how a chromosome domain with essential function

can evolve so rapidly[10,13]. Support for this hypothesis was originally based on centromere sequence divergence between more distantly related species and the rapid evolution of centromeric proteins[10,13]. Our study highlights how rapid this centromere sequence evolution can be. We hypothesize that all observations described in the present study are signatures of genetic conflict. Based on this hypothesis, we build an evolutionary model that summarizes the rapid turnover of centromeric sequences in the *melanogaster* clade species (Figure 7B).

We propose that the ancestor of *D. melanogaster* and the *D. simulans* clade had retroelement-rich centromeres. Then, *500-bp* and *365-bp* complex satellites invaded the centromeres in the common ancestor of the *D. simulans* clade species. These selfish centromeric repeats replaced the ancestral retroelement-rich centromeres. The *500-bp* satellite is also abundant in pericentromeres (Figures 1 and 2). The spread of *500-bp* to what is now pericentromeres may be a signature of past invasion – CENP-A may have restricted its domain to a subset of the *500-bp* satellite array to avoid centromere asymmetry. The only retroelement countering this invasion, and preventing the complete homogenization of centromeres, is *G2/Jockey3.* This retroelement seems to have actively targeted centromeric chromatin in the ancestor of *D. melanogaster* and the *D. simulans* clade and continued its activity in *D. melanogaster* [62] and *D. simulans*. Thus, while all other centromeric sequences were replaced, *G2/Jockey3* continued jumping into the centromere chromatin, until it lost its ability to mobilize and was gradually lost in *D. mauritania* and *D. sechellia*. Following the invasion of the complex satellites, the centromere shifted to the neighboring telomeric HTT in *D. sechellia*, replacing the centromeric satellites on the X and dot chromosomes. We speculate that their telomeric function may give them some transmission advantage, but further experiments are required to test this hypothesis. It will be difficult to gather direct empirical evidence for selfish dynamics in the turnover after sequences are stabilized at centromeres. However, indirect support for the selfish turnover hypothesis comes from the comparative analysis of Y chromosome centromeres.

Drosophila Y chromosome centromeres have an unusual composition: the ~40-Mb Y chromosomes of *D. melanogaster*[40] and the *D. simulans* clade[30] are extremely rich in simple satellite sequences (up to 80% of the Y chromosome[39]), yet their Y centromeres are enriched in TEs and completely devoid of satellite sequences[25] (Figure 5). Selfish centromeres should only

have an advantage in female meiosis where there is an asymmetry: only one of the products of meiosis contributes to the next generation. The unique composition of Y chromosome centromeres may be because it is the only chromosome that never experiences female meiosis and is thus not susceptible to selfish centromere invasion (Figure 7B). While selfish centromere drivers (*e.g.,* driving satellites) cannot invade Y chromosomes, these chromosomes still offer a safe haven for transposable element insertions.

In conclusion, we speculate that the extremely rapid turnover of centromeric DNA in the *D. melanogaster* subgroup is driven by selfish dynamics. If the genetic elements occupying centromeres are selfish, competition for centromere invasion and potential for biased transmission to the next generation can drive rapid turnover of centromere composition. In these species, retroelements and satellite DNA may be competing for centromere occupancy. These dynamics have implications not just for the role of centromeric DNAs in chromosome segregation, but also for the role of retroelements in genome function, and karyotype evolution broadly.

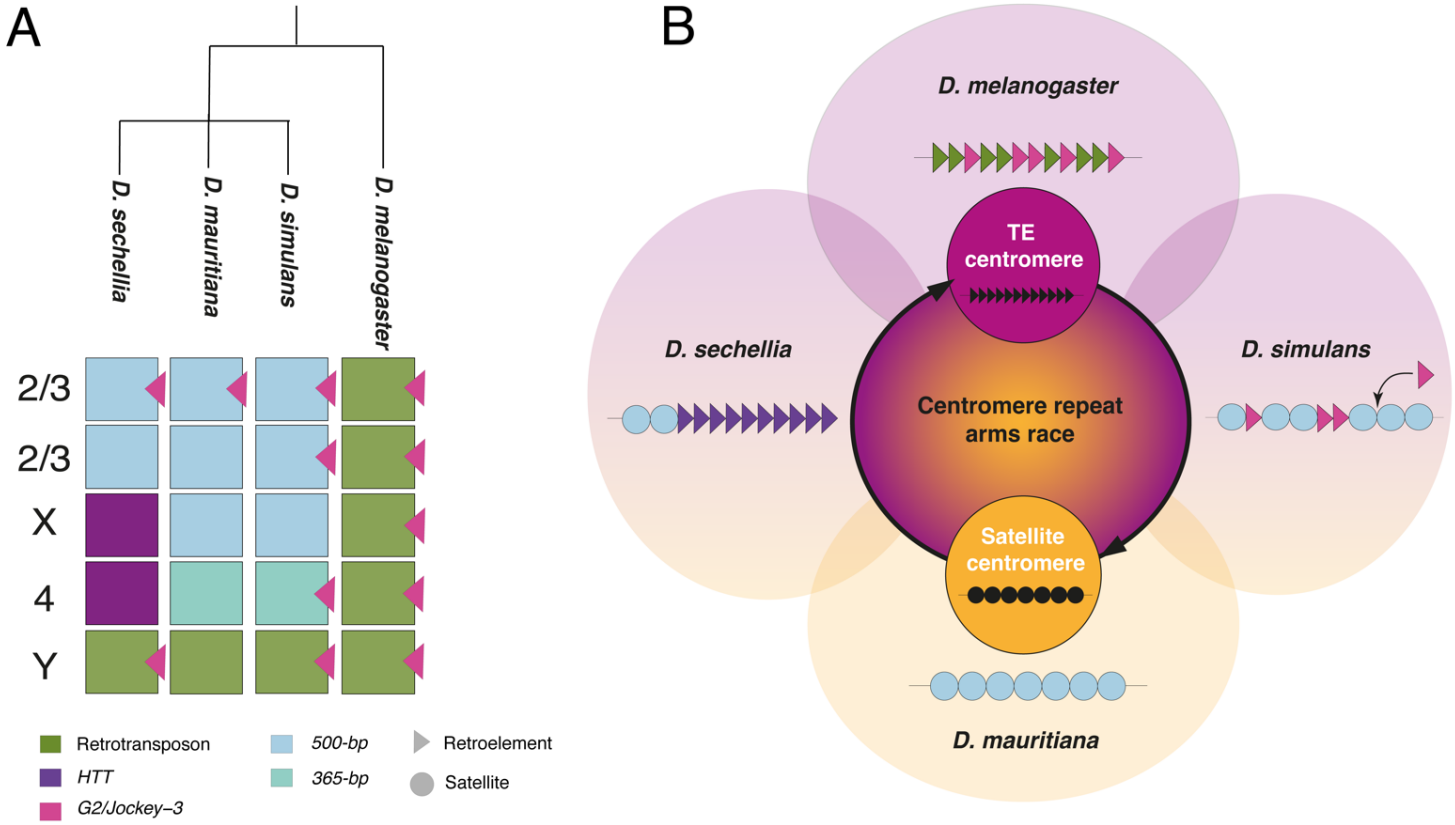


Figure 7. Shifting centromere composition in the *D. simulans* clade species and *D. melanogaster*. A) A summary of the turnover in centromere composition between species where each square represents a centromere, colored based on the most abundant repeat. The color code is indicated in the legend. Centromeres containing at least one *G2/Jockey3* insertion are marked by a magenta triangle. **B)** An evolutionary model for the centromere sequence turnover in the *melanogaster* clade species: we propose an arms race (inner circle) drove the turnover between retroelement-rich and satellite-rich centromeres in the simulans clade species. Representative examples of specific replacement events in different stages of the arms race are depicted in the outside circles.

MATERIALS AND METHODS

Fly strains

For *D. sechellia* and *D. mauritiana,* we used the same sequenced strains used to build the heterochromatin enriched genome assemblies[30]: Rob12 (Cornell SKU: 14021-0248.25) and w12 (Cornell SKU :14021-0241.151), respectively. For *D. simulans,* we used the wXD1 strain that is maintained in the Larracuente lab. While it is the same strain as the one used to build the heterochromatin enriched assembly, our isolate appears to have a structural polymorphism on the X chromosome pericentromeric compared to the assembly[30]. All the experiments conducted in this study were performed using the same isolate. For *D. melanogaster,* we used an inbred strain from the Netherlands (N25)[69].

Antibodies used

The list of primary and secondary antibodies that we used for this study is details below :

- anti-CENP-A antibody ($\boldsymbol{\alpha}$-CID20): polyclonal rabbit antibody synthesis for this study (by Covalance). The CENP-A antibody was raised against the MPRHSRAKRAPRPSAC peptide[26]. The final serum was proteinA purified. We used this antibody at 1:50 dilution for the CUT&Tag.
- anti-CENP-C antibody ($\boldsymbol{\alpha}$-CENP-C12): polyclonal rabbit antibody synthesis for this study (by Genescript). The CENP-C antibody was raised against the NNRRSMRRSGNVPGC peptide. The final serum was affinity purified. We used this antibody at 1:100 dilution for the Immunostaining on mitotic chromosomes.
- anti-CENP-A antibody ($\boldsymbol{\alpha}$-CIDH32) : polyclonal chicken antibody, gift from the Mellone lab. We used the antibody at 1:100 dilution for the Immunostaining on chromatin fibers.
- Anti-Mouse IgG H&L antibody (abcam, ab46540): rabbit antibody that we used as a negative control for the CUT&Tag at 1:100 dilution.
- anti-H3K9me3 antibody (abcam, ab176916): rabbit monoclonal antibody. We used this antibody as a positive control for the CUT&Tag at 1:100 dilution.

- anti CENP-C primary antibody : Guinea Pig antibody from[70]. We used this antibody for larval brain squashes for *G2/Jockey-3* IF-FISH at 1:500 dilution.
- Guinea Pig anti-rabbit unconjugated (Novus Biologicals, NBP1-72763). We used this secondary antibody for the CUT&Tag at 1:100 dilution.
- Goat anti-rabbit IgI H&L conjugate with Alexa Fluor 488 (abcam, ab150077). We used this secondary antibody for the Immunostaining on mitotic chromosomes spread at 1:500 dilution.
- Goat anti-Chicken IgY (H+L) Secondary Antibody, Alexa Fluor™ 488 (Invitrogen, A-11039)
- Goat anti Guinea Pig conjugate with AlexaFlour 546 (Thermo Catalog # A-11074). We used this secondary antibody for the Immunostaining on mitotic chromosomes spread at 1:500, for *G2/Jockey-3* IF-FISH.

CUT&Tag

Nuclei isolation

We collected Drosophila embryos overnight at 25°C in cages containing a grape juice-agar plate with yeast paste. We used 0-16h embryos to perform nuclei isolation as in[71]. We washed embryos in the embryo wash buffer (0.7% NaCl, 0.04% Triton-X100) then dechorionated using 50% bleach for 30s. We ground embryos in 1ml buffer B (pH7.5, 15mM Tris-HCl, 15mM NaCl, 60mM KCl, 0.34M Sucrose, 0.5mM Spermidine, 0.1% β-mercaptoethanol, 0.25mM PMSF, 2mM EDTA, 0.5mM EGTA) using a homogenizer and filtered to remove large debris. We centrifuged nuclei at 5000g for 5 min and resuspended in 500µl of buffer A (pH7.5, 15mM Tris-HCl, 15mM NaCl, 60mM KCl, 0.34M Sucrose, 0.5mM Spermidine, 0.1% β-mercaptoethanol, 0.25mM PMSF), twice. We resuspended the final pellet in CUT&Tag wash buffer (20mM HEPES pH 7.5, 150mM NaCl, 0.5 mM Spermidine) to a final concentration of 1 000 000 nuclei/ml.

CUT&Tag

We performed CUT&Tag using around 100,000 nuclei per sample. We used the pA-Tn5 enzyme from Epycpher and followed the manufacturer's protocol (CUT&Tag Protocol v1.5). For each species we performed 3 replicates with the anti-CID20 antibody (1:50), one positive control using anti-H3K9me3 (1:100), and one negative control using the anti-IgG antibody (1:100).

Library preparation
For the library preparation, we used the primers from[72] (SupTable7). Before final sequencing, we pooled 2ul of each library and performed a MiSeq run. We used the number of resulting reads from each library to estimate the relative concentration of each library and ensure an equal representation of each library in the final pool for sequencing. We sequenced the libraries in 150-bp paired-end mode on HiSeq Illumina. We obtained around 10 million reads per library, except for the IgG negative control that usually has a lower representation (SupTable8).

Centromere identification

We trimmed paired-end reads using trimgalore (v0.4.4) (*trim_galore --paired --nextera --length 75 --phred33 --no_report_file –fastqc*)[73] and assessed read quality with FASTQC.

We mapped reads against the reference genome with bwa (v7.4) using the *BWA-MEM* algorithm (default parameters). We used the heterochromatin-enriched assemblies of *D. melanogaster* [40], *D. simulans, D. sechellia* and *D. mauritiana*[30]. We converted the resulting sam alignment files into bam files and sorted using respectively samtools (v1.11) *view* and *sort* command. We removed PCR duplicates using *Markduplicates* from Picardtools (v2.12.0) (https://broadinstitute.github.io/picard/). Because we are working with highly repetitive sequences we analyzed both the unique and mutli-mapping reads. We thus performed two different filtering based on mapping quality using samtools *view*[74]. To keep multi-mapping reads, we use the following parameters: *-b -h -f 3 -F 4 -F 8 -F 256 -F 2048*. To keep only the uniquely mapping reads we use the following parameters: *-b -h -f 3 -F 4 -F 8 -F 256 -F 2048 -q30.*
We estimated read coverage using the *bamCoverage* command from deeptools (v3.5.1) using the option *--scaleFactor -bs 1 --extendReads* and normalized the read coverage to RPM (reads per million).
We called peaks using MACS2 callpeak (v2017-10-26) (option -f BAMPE -g dm -q 0.01 -B --call-summits) and performed an IDR analysis (https://github.com/nboley/idr) to identify high confidence peaks that overlapped between replicates (IDR <0.05, SupTable1). The localization of these high confident peaks allowed us to identify the candidate centromere contigs (SupFig1).

Repeat enrichment analysis

For this analysis, we used the multi-mapping bam file. We annotated the reference genome (SupFile1-4) using a custom repeat library specific to each species (SupFile5-8) with Repeatmasker (options *-no_is -a -inv -pa 20 -div 20*)[75]. Using htseq-count[76] we counted the number of reads that map to each repeat and calculated RPM. To determine the enrichment, we normalized the RPM counts for CID by RPM counts for IgG (negative control). The 25 % most enriched repeats are presented in SupTable9, and the top 20 most enriched repeats among all replicates are presented in Figure 1 B, E, H.

To explore origins of the centromeric complex satellites we blasted (*blastn* with default parameter) the consensus sequences of *500-bp, 136-bp* and *365-bp* satellites against the genome of *D. melanogaster*[40]*,* the *simulans* clade[30] and more distant species*, D. yakuba, D. annanassae, D. pseudoobscura, D. erecta* and *D. virilis*[77]. All hits are reported in SupTable3-5.

The dotplots of the Y chromosome centromeres cenY (SupFig9) were generated using re-DOT-able v1.1 (https://www.bioinformatics.babraham.ac.uk/projects/redotable/).

G2/Jockey3 evolutionary analyses

We surveyed *G2/Jockey-3* evolution in additional species with improved genome assemblies of *D. simulans*, *D. sechellia*, and *D. mauritania*[78] and publicly-available Nanopore assemblies of *D. yakuba*, *D. erecta*, and *D. ananassae*[79]. We identified *G2/Jockey-3* sequences with two complementary methods. First, we annotated each genome assembly with our custom Drosophila TE library including the *D. melanogaster G2/Jockey-3* consensus sequence[62] using Repeatmasker v4.1.0. The annotations and 500 bp flanking regions were extracted with BEDTools v2.29.0[80] and aligned with MAFFT[81] to generate a species-specific consensus sequence with Geneious v.8.1.6 [82]. Each assembly was annotated again using Repeatmasker with the appropriate species-specific *G2/Jockey-3* consensus sequence. Second, we constructed *de novo* repeat libraries for each species with RepeatModeler2 v.2.0.1[83] and identified candidate *G2/Jockey-3* sequences which shared high similarity with *G2/Jockey-3* in *D. melanogaster* identified with BLAST v.2.10.0. We did the same with *Jockey-1* (LINEJ1_DM) as method confirmation and to use as an outgroup for TE fragment alignment. We removed candidates shorter than 100 bp from the analysis. We identified ORFs within consensus TE sequences with NCBI ORFfinder. We used Repeatmasker to annotate the genome assemblies with the *de novo Jockey-3* consensus sequences. To infer a phylogenetic tree

of TEs, we aligned *G2/Jockey-3* fragments identified in each species with MAFFT and retained sequences corresponding to the ORF bounds of the consensus sequences; We removed ORF fragments <400 bp. We inferred the tree with RAxML v.8.2.11[84] using the command "raxmlHPC-PTHREADS -s alignment_Jockey-3_melsimyak_400_ORF2_mafft.fasta -m GTRGAMMA -T 24 -d -p 12345 -# autoMRE -k -x 12345 -f a".

Oligopaint design and synthesis

We designed oligopaints probes targeting *500-bp, 136-bp, 365-bp, Rsp-like*, *HTTs* and the Y centromere islands of each species using ProbeDealer[85] with some modifications. We extracted the fasta sequences containing the target repeat from the reference genomes and used it as the input for ProbeDealer. After designing all the possible oligo probes, ProbeDealer usually maps them back against the reference genome to eliminate multimapping oligos. Because we are working with highly repetitive sequences, we skipped this step. We mapped the oligos to the reference genome to manually inspect for potential off targets. The final oligo list is in SupTable10. Oligopaints libraries were synthesized by Genescript. We then synthesized and purified each oligopaint sublibrary as described in[25].

IF-FISH on mitotic chromosome

We dissected brains from third instar larvae in PBS, incubated 8 min in 0.5% sodium citrate. We fixed for 6 min in 4% formaldehyde, 45% acetic acid before squashing. We squashed the brains between a poly-lysine slide and coverslip and before immersing in liquid nitrogen. After 5 min in PBS and 10 min in PBS, we blocked slides for at least 30 min in blocking buffer (3%BSA, 1% goat serum in PBST). For immunofluorescence (IF), we incubated slides in primary antibody (***α***-CENP-C12 1:100) overnight at 4°C. We washed slides 3 times for 5 min in PBST. We incubated slides in secondary antibody (anti-rabbit 1:500) for 1-3h at room temperature and washed 3 times for 5 min in PBST. We post-fixed the slides using 10% formaldehyde diluted in 4XSSC, incubating 20 min at room temperature and washed 3 times for 3 min with 4XSSC and one time for 5 min with 2XSSC. For the hybridization, we used 20 pmol of primary probes (SupTableOligo) and 80 pmol of the secondary probes (SupTable11) in 50 µl of hybridization buffer (50% formamide, 10% dextran sulfate, 2XSSC). We heated slides for 5 min at 95°C to denature and incubated them

overnight at 37°C in a humid chamber. We then washed the slides 3 times for 5 min with 4XSSCT and 3 times for 5min with 0.1SSC before mounting in slowfade DAPI.

IF-FISH on chromatin fibers

We dissected 3rd instar larval brains in 1XPBS (3-4 brains per slide) and incubated in 250µl of 0.5% sodium citrate with 40µg of dispase-collagenase, for 12 min at 37°C. The tissue was transferred to a poly-lysine slide using Shandon Cytospin 4 at 1,200 rpm for 5 minutes. We positioned slides vertically in a tube containing the Lysis buffer (500nM NaCl, 25mM Tris-HCL pH7.5, 250nM Urea, 1% Triton X-100) and incubated for 16 min. For the fiber stretching, we allow the buffer to slowly drain from the tube with the hole at the bottom (by removing the tape). A steady flow rate will generate a hydrodynamic drag force which generates longer and straighter fibers. We incubated slides in a fixative buffer (4% formaldehyde) for 10 min and then 10 min in 1XPBST (0.1% Triton). For the IF, we first blocked the slides for 30 min in blocking buffer (1.5% BSA in 1XPBS). We incubated slides overnight at 4°C with the primary antibody (***α***-CIDH32, 1:100) and washed 3 times for 5 min in 1xPBST. We incubated slides with the secondary antibody (anti-chicken, 1:500) for 1-3 h at room temperature and washed 3 times for 5min with 1XPBST. We post-fixed the slide with 10% formaldehyde for 20 min and washed 3 times for 5 min in 1XPBST. We then incubated slides for 10 min in 2XSSCT at room temperature and 10 min in 2XSSCT - 50% formamide at 60°C. For the hybridization, we used 40 pmol of primary probes (SupTableOligo) and 160 pmol of the secondary probes (SupTableProbes) in 100 µl of hybridization buffer (50% formamide, 10% dextran sulfate, 2XSSC). We heated slides for 5 min at 95°C to denature and incubated them overnight at 37°C in a humid chamber. We then washed the slides 15 min with 2XSSCT at 60°C, 15 min with 2XSSCT at room temperature, and 10 min with 0.1XSSC at room temperature. We incubated slides for 5 min in DAPI (1mg/ml) before mounting in SlowFadeTM Gold (Invitrogen S36936).

G2/Jockey-3 IF-FISH

D. simulans, *D. sechellia,* and *D. mauritania* third instar larval brains were dissected in 1X PBS and all attached tissue or mouth parts were removed with forceps. Brains were immersed in 0.5% sodium citrate solution for 8 min in a spot well dish. The tissue was placed in a 6µl drop of 45% acetic acid, 2% Formaldehyde on a siliconized (Rain X) coverslip for 6 min. A poly-lysine coated

slide was inverted and placed on the brains to make a sandwich. After flipping the slide and gently removing excess fixative between a bibulous paper, the brain was squashed using the thumb by firmly pressing down. Slides were then immersed in liquid nitrogen and the coverslip flipped off using a razor blade and transferred to 1X PBS for 5 min to rehydrate before proceeding with IF-FISH. Slides were then washed with 1X PBST(0.1% Triton X-100) for 5 min on a rotator, repeated 3 times. Slides were then transferred to a coplin jar containing blocking solution (1% BSA in 1X PBST) for 30 min while rocking. Diluted antibodies were applied to the slides coating the brains with 50 µl of primary antibodies, covered with parafilm and stored in a dark chamber at 4oC overnight. The following day slides were washed 4 times with 1X PBST for 5 min while rocking. Secondary antibodies diluted with block were applied to the brains and covered with paraflim, then incubated at room temperature for 1 hr. After the 1 hr incubation, slides were washed 4 times in 1X PBST for 5 min while rotating. Slides were then post-fixed with 3.7% Formaldehyde diluted with 1X PBS for 10 min in the dark. Slides were washed for 5 min in 1X PBS while rotating before proceeding to FISH. The following FISH protocol for G2/Jockey-3 labeling and the synthesis of the *G2/Jockey-3* probe was performed as described in the methods of Chang et al 2019. Slides were dehydrated in an ethanol row (3 min washes in 70%, 90%, and 100% ethanol) and allowed to air-dry completely for a few minutes. Probe mix (20 µL) containing 2xSSC, 50% formamide (Sigma-Aldrich), 10% dextran sulfate (Merck), 1 µL RNase cocktail (ThermoFisher), and 100 ng of DIG-labeled *G2/Jockey-3* probe was boiled at 80°C for 8 min, incubated on ice for 5 min, and then applied to slides, covered with a glass coverslip, and sealed with paper cement. Sealed slides were denatured on a slide thermocycler for 5 min at 95°C and incubated at 37°C overnight to hybridize. Slides were then washed three times in a coplin jar for 5 min in 2xSSC, 50% formamide at 42°C. Slides were then washed three times for 5 min in 0.1xSSC at 60°C, and then blocked in block buffer 1% BSA, 4xSSC, 0.1% Tween-20 at 37°C for 45 min. Slides were incubated with 50 µL of block buffer containing a fluorescein-labeled anti-DIG antibody (sheep, 1:100, Roche) for 60 min at 37°C. Slides were then washed three times for 5 min in 4xSSC, 0.1% Tween-20 at 42°C. Slides were washed with 1X PBS briefly in a coplin jar and finally mounted on a coverslip with Slowfade and DAPI, then sealed with nail polish.

Image acquisition

We imaged using a LEICA DM5500 microscope with a 100x/oil immersion objective or Delta vision using an Olympus UPLansApo 100x/1.40 oil immersion objective, maintaining all exposures consistent across each experiment. Images obtained with the Deltavision microscope were deconvolved with Softoworks using 5 iterations with the 'conservative' setting. Images were edited, cropped and pseudocolored using Fiji.

Data availability

All sequences are available from NCBI SRA under Bioproject accession PRJNA1007690 All the BASH pipelines and R scripts used in this study are available on github: https://github.com/LarracuenteLab/SimClade_Centromere_2023. All files necessary to reproduce the plots are on Dryad (DOI forthcoming).

Acknowledgment

We would like to thank the members of the Larracuente and Mellone labs for helpful discussion and Emiliano Martí for comments on the manuscript. This research was supported by the National Science Foundation (Grant number MCB-1844693). We are grateful to the University of Rochester Center for Integrated Research Computing for access to computing cluster resources and the University of Rochester Genomics Research Center for sequencing services.

SUPPLEMENTAL TABLES

SupTable 1: Output of the IDR analyses for both the uniquely (Q30) and multimapping (Q0) peak calling for each species.

SupTable 2 : Coordinate of each centromere of the *melanogaster* clade.

SupTable 3 : Output of the blast for *500-bp*.

SupTable 4 : Output of the blast for *136-bp*.

SupTable 5 : Output of the blast for *365-bp*.

SupTable 6 : Composition of the Y chromosome centromeres.

SupTable 7 : List of the primers used for the CUT&Tag libraries.

SupTable 8 : Number of reads and percentage of mapping for each CUT&Tag libraries.

SupTable 9 : Repeat enrichment in RPM

SupTable 10 : Oligopaints sequences

SupTable 11 : Sequences of probes used for FISH.

SUPPLEMENTAL FILES

SupFile1 : *D. simulans* genome annotation

SupFile2 : *D. mauritiana* genome annotation

SupFile3 : *D. sechellia* genome annotation

SupFile4 : *D. melanogaster* genome annotation

SupFile5 : *D. simulans* repeat library

SupFile6 : *D. mauritiana* repeat library

SupFile7 : *D. sechellia* repeat library

SupFile8 : *D. melanogaster* repeat library

REFERENCES

1. Barra, V. & Fachinetti, D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat. Commun.* **9**, 4340 (2018).
2. Karpen, G. H. & Allshire, R. C. The case for epigenetic effects on centromere identity and function. *Trends Genet.* **13**, 489–496 (1997).
3. Mendiburo, M. J., Padeken, J., Fülöp, S., Schepers, A. & Heun, P. Drosophila CENH3 is sufficient for centromere formation. *Science* **334**, 686–690 (2011).
4. Mellone, B. G. & Fachinetti, D. Diverse mechanisms of centromere specification. *Curr. Biol.* **31**, R1491–R1504 (2021).
5. Sun, X. *et al.* Elevated expression of the centromere protein-A(CENP-A)-encoding gene as a prognostic and predictive biomarker in human cancers: Elevated Expression of the CENP-A-Encoding Gene in Cancer. *Int. J. Cancer* **139**, 899–907 (2016).
6. Talbert, P. B. & Henikoff, S. What makes a centromere? *Exp. Cell Res.* **389**, 111895 (2020).
7. Saha, A. K. *et al.* The Genomic Landscape of Centromeres in Cancers. *Sci. Rep.* **9**, 11259 (2019).
8. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* **9**, 615 (2018).
9. Melters, D. P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
10. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
11. Haaf, T. & Willard, H. F. Chromosome-specific α-satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma* **106**, 226–232 (1997).
12. Haaf, T. & Willard, H. F. Orangutan α-satellite monomers are closely related to the human consensus sequence. *Mamm. Genome* **9**, 440–447 (1998).
13. Malik, H. S. & Henikoff, S. Major Evolutionary Transitions in Centromere Complexity. *Cell* **138**, 1067–1082 (2009).
14. Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**, 1559–1562 (2008).
15. De Villena, F. P.-M. & Sapienza, C. Female Meiosis Drives Karyotypic Evolution in Mammals. *Genetics* **159**, 1179–1189 (2001).
16. Chmátal, L. *et al.* Centromere Strength Provides the Cell Biological Basis for Meiotic Drive and Karyotype Evolution in Mice. *Curr. Biol.* **24**, 2295–2300 (2014).
17. Akera, T. *et al.* Spindle asymmetry drives non-Mendelian chromosome segregation. *Science* **358**, 668–672 (2017).
18. Iwata-Otsubo, A. *et al.* Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. *Curr. Biol. CB* **27**, 2365-2373.e8 (2017).
19. Presting, G. G. Centromeric retrotransposons and centromere function. *Curr. Opin. Genet. Dev.* **49**, 79–84 (2018).
20. Hickey, D. A. SELFISH DNA: A SEXUALLY-TRANSMITTED NUCLEAR PARASITE. *Genetics* **101**, 519–531 (1982).
21. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
22. Patchigolla, V. S. P. & Mellone, B. G. Enrichment of Non-B-Form DNA at *D.*

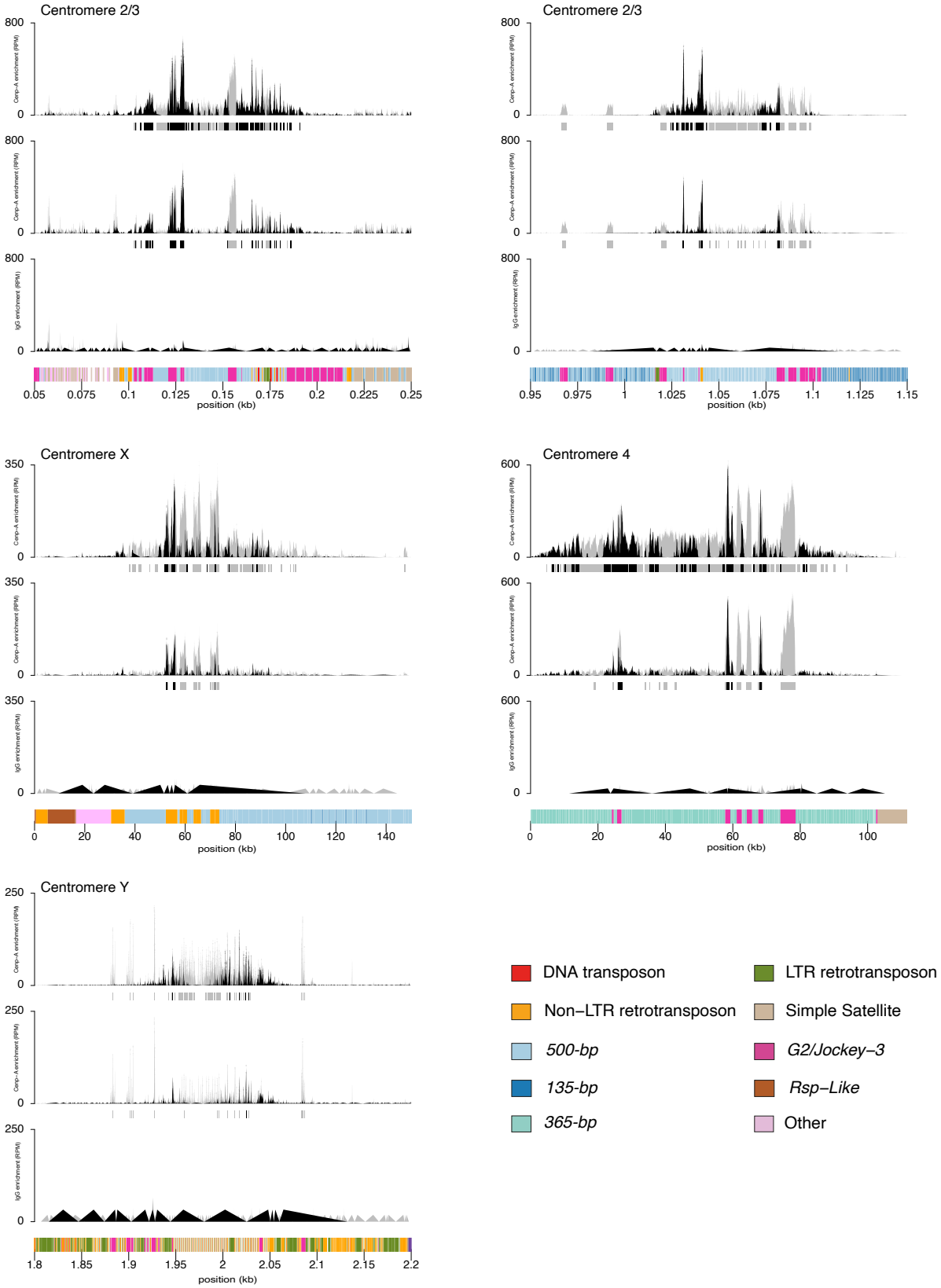
melanogaster Centromeres. *Genome Biol. Evol.* **14**, evac054 (2022).
23. Csink, A. K. & Henikoff, S. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204 (1998).
24. Klein, S. J. & O'Neill, R. J. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* **26**, 5–23 (2018).
25. Chang, C.-H. *et al.* Islands of retroelements are major components of Drosophila centromeres. *PLOS Biol.* **17**, e3000241 (2019).
26. Talbert, P. B., Kasinathan, S. & Henikoff, S. Simple and Complex Centromeric Satellites in Drosophila Sibling Species. *Genetics* **208**, 977–990 (2018).
27. Lachaise, D. *et al.* Historical Biogeography of the *Drosophila melanogaster* Species Subgroup. in *Evolutionary Biology* (eds. Hecht, M. K., Wallace, B. & Prance, G. T.) 159–225 (Springer US, 1988). doi:10.1007/978-1-4613-0931-4_4.
28. Russo. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* (1995) doi:10.1093/oxfordjournals.molbev.a040214.
29. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
30. Chang, C.-H., Gregory, L. E., Gordon, K. E., Meiklejohn, C. D. & Larracuente, A. M. Unique structure and positive selection promote the rapid divergence of Drosophila Y chromosomes. *eLife* **11**, e75795 (2022).
31. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).
32. Beliveau, B. J., Apostolopoulos, N. & Wu, C. Visualizing Genomes with Oligopaint FISH Probes. *Curr. Protoc. Mol. Biol.* **105**, (2014).
33. Larracuente, A. M. The organization and evolution of the *Responder* satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. *BMC Evol. Biol.* **14**, 233 (2014).
34. Sproul, J. S. *et al.* Dynamic Evolution of Euchromatic Satellites on the X Chromosome in *Drosophila melanogaster* and the simulans Clade. *Mol. Biol. Evol.* **37**, 2241–2256 (2020).
35. Pardue, M. L., Danilevskaya, O. N., Lowenhaupt, K., Slot, F. & Traverse, K. L. Drosophila telomeres: new views on chromosome evolution. *Trends Genet.* **12**, 48–52 (1996).
36. Jagannathan, M., Warsinger-Pepe, N., Watase, G. J. & Yamashita, Y. M. Comparative Analysis of Satellite DNA in the *Drosophila melanogaster* Species Complex. *G3 GenesGenomesGenetics* **7**, 693–704 (2017).
37. Kaufman, T. C. A Short History and Description of *Drosophila melanogaster* Classical Genetics: Chromosome Aberrations, Forward Genetic Screens, and the Nature of Mutations. *Genetics* **206**, 665–689 (2017).
38. Marchetti, M. *et al.* Cytological heterogeneity of heterochromatin among 10 sequenced *Drosophila* species. *Genetics* **222**, iyac119 (2022).
39. Lohe, A. R., Hilliker, A. J. & Roberts, P. A. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics* **134**, 1149–1174 (1993).
40. Chang, C.-H. & Larracuente, A. M. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster* Y Chromosome. *Genetics* **211**, 333–348 (2019).
41. Tambones, I. L., Haudry, A., Simão, M. C. & Carareto, C. M. A. High frequency of horizontal transfer in Jockey families (LINE order) of drosophilids. *Mob. DNA* **10**, 43 (2019).
42. Wang, N., Liu, J., Ricci, W. A., Gent, J. I. & Dawe, R. K. Maize centromeric chromatin scales with changes in genome size. *Genetics* **217**, iyab020 (2021).

43. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
44. Logsdon, G. A. *et al. The variation and evolution of complete human centromeres*. http://biorxiv.org/lookup/doi/10.1101/2023.05.30.542849 (2023) doi:10.1101/2023.05.30.542849.
45. Wlodzimierz, P. *et al.* Cycles of satellite and transposon evolution in Arabidopsis centromeres. *Nature* **618**, 557–565 (2023).
46. Singchat, W. *et al.* Population Scale Analysis of Centromeric Satellite DNA Reveals Highly Dynamic Evolutionary Patterns and Genomic Organization in Long-Tailed and Rhesus Macaques. *Cells* **11**, 1953 (2022).
47. Suzuki, Y., Myers, E. W. & Morishita, S. Rapid and ongoing evolution of repetitive sequence structures in human centromeres. *Sci. Adv.* **6**, eabd9230 (2020).
48. Arora, U. P., Charlebois, C., Lawal, R. A. & Dumont, B. L. Population and subspecies diversity at mouse centromere satellites. *BMC Genomics* **22**, 279 (2021).
49. Yang, Z. *et al.* Cotton D genome assemblies built with long-read data unveil mechanisms of centromere evolution and stress tolerance divergence. *BMC Biol.* **19**, 115 (2021).
50. Koch, M. A. & Kiefer, M. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species--*Capsella rubella*, *Arabidopsis lyrata* subsp. petraea, and *A. thaliana*. *Am. J. Bot.* **92**, 761–767 (2005).
51. Kipling, D., Ackford, H. E., Taylor, B. A. & Cooke, H. J. Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics* **11**, 235–241 (1991).
52. Garagna, S., Zuccotti, M., Capanna, E. & Redi, C. A. High-resolution organization of mouse telomeric and pericentromeric DNA. *Cytogenet. Genome Res.* **96**, 125–129 (2002).
53. Fennell, A., Fernández-Álvarez, A., Tomita, K. & Cooper, J. P. Telomeres and centromeres have interchangeable roles in promoting meiotic spindle formation. *J. Cell Biol.* **208**, 415–428 (2015).
54. Jeon, H.-J. & Oh, J. S. TRF1 Depletion Reveals Mutual Regulation Between Telomeres, Kinetochores, and Inner Centromeres in Mouse Oocytes. *Front. Cell Dev. Biol.* **9**, 749116 (2021).
55. Birchler, J. A. & Presting, G. G. Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes. *Genes Dev.* **26**, 638–640 (2012).
56. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308 (2017).
57. Bergmann, J. H. *et al.* Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore: H3K4me2 and kinetochore maintenance. *EMBO J.* **30**, 328–340 (2011).
58. Carone, D. M. *et al.* Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. *Chromosome Res.* **21**, 49–62 (2013).
59. Chen, C.-C. *et al.* Establishment of Centromeric Chromatin by the CENP-A Assembly Factor CAL1 Requires FACT-Mediated Transcription. *Dev. Cell* **34**, 73–84 (2015).
60. Mejía, J. E. *et al.* Efficiency of de Novo Centromere Formation in Human Artificial Chromosomes. *Genomics* **79**, 297–304 (2002).
61. Quénet, D. & Dalal, Y. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. *eLife* **3**, e26016 (2014).
62. Hemmer, L. W. *et al.* Centromere-associated retroelement evolution in Drosophila melanogaster reveals an underlying conflict.

http://biorxiv.org/lookup/doi/10.1101/2022.11.25.518008 (2022) doi:10.1101/2022.11.25.518008.
63. Palladino, J., Chavan, A., Sposato, A., Mason, T. D. & Mellone, B. G. Targeted De Novo Centromere Formation in Drosophila Reveals Plasticity and Maintenance Potential of CENP-A Chromatin. *Dev. Cell* **52**, 379-394.e7 (2020).
64. Wolfgruber, T. K. *et al.* Maize Centromere Structure and Evolution: Sequence Analysis of Centromeres 2 and 5 Reveals Dynamic Loci Shaped Primarily by Retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
65. Schneider, K. L., Xie, Z., Wolfgruber, T. K. & Presting, G. G. Inbreeding drives maize centromere evolution. *Proc. Natl. Acad. Sci.* **113**, (2016).
66. Tsukahara, S. *et al.* Bursts of retrotransposition reproduced in Arabidopsis. *Nature* **461**, 423–426 (2009).
67. Tsukahara, S. *et al.* Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes Dev.* **26**, 705–713 (2012).
68. Kawabe, A. & Nasuda, S. Structure and genomic organization of centromeric repeats in Arabidopsis species. *Mol. Genet. Genomics* **272**, 593–602 (2005).
69. Grenier, J. K. *et al.* Global Diversity Lines–A Five-Continent Reference Panel of Sequenced *Drosophila melanogaster* Strains. *G3 GenesGenomesGenetics* **5**, 593–603 (2015).
70. Erhardt, S. *et al.* Genome-wide analysis reveals a cell cycle–dependent mechanism controlling centromere propagation. *J. Cell Biol.* **183**, 805–818 (2008).
71. Chen, T. *et al.* The nanoCUT&RUN technique visualizes telomeric chromatin in Drosophila. *PLOS Genet.* **18**, e1010351 (2022).
72. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, (2015).
73. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo. (2021) doi:10.5281/ZENODO.5127899.
74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
75. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. (2013).
76. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945 (2022).
77. Kim, B. Y. *et al.* Highly contiguous assemblies of 101 drosophilid genomes. *eLife* **10**, e66405 (2021).
78. Chakraborty, M. *et al.* Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res.* **31**, 380–396 (2021).
79. Miller, D. E., Staber, C., Zeitlinger, J. & Hawley, R. S. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3 GenesGenomesGenetics* **8**, 3131–3141 (2018).
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
82. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
83. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
84. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of

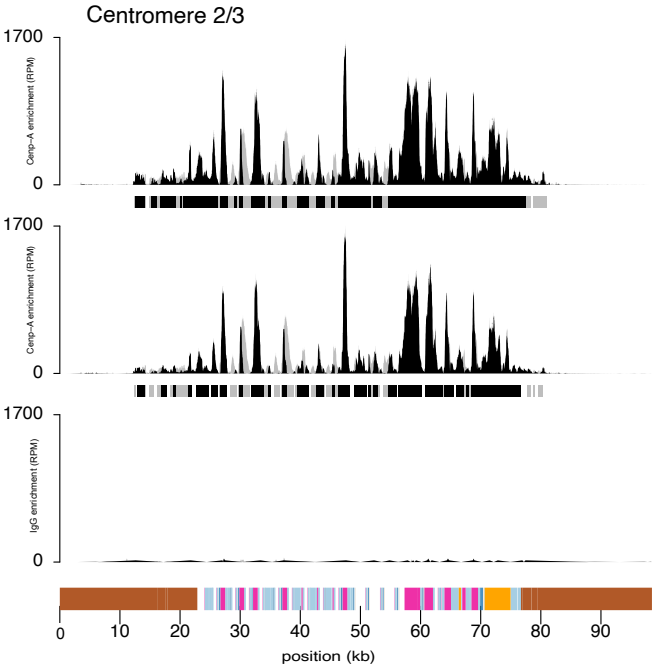
large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
85. Hu, M. *et al.* ProbeDealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization. *Sci. Rep.* **10**, 22031 (2020).

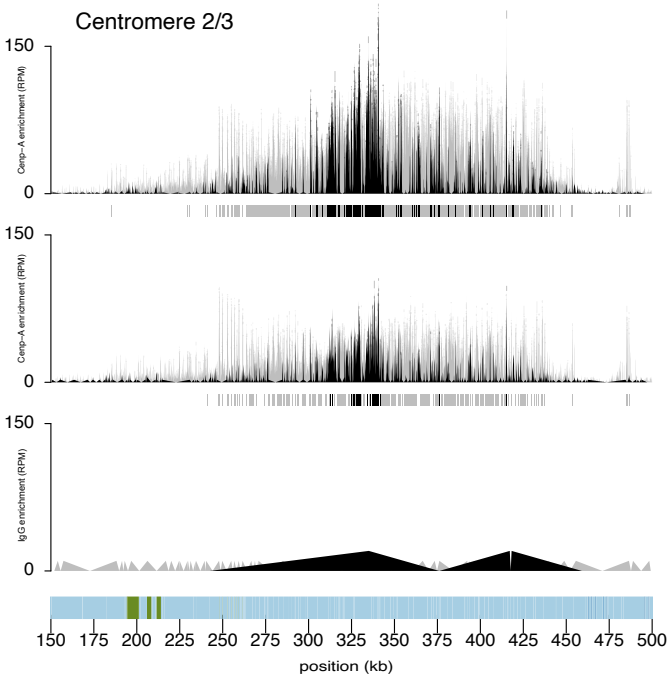
EXTENDED DATA

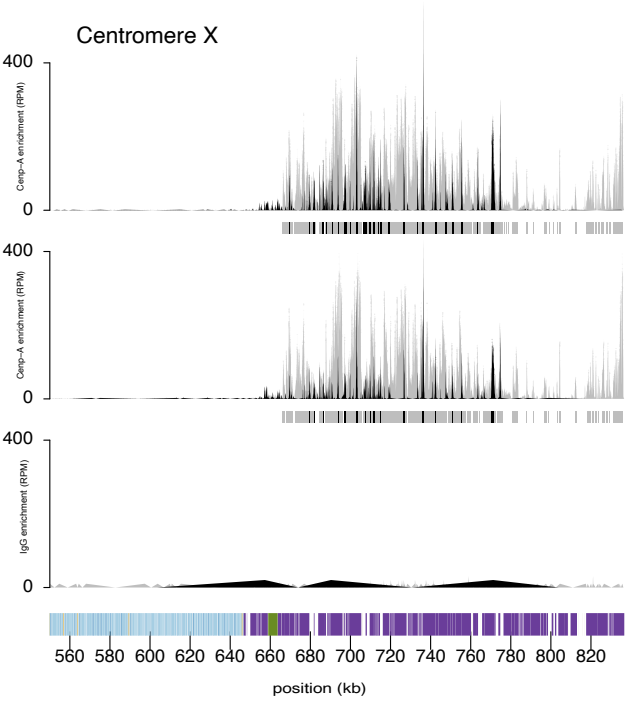


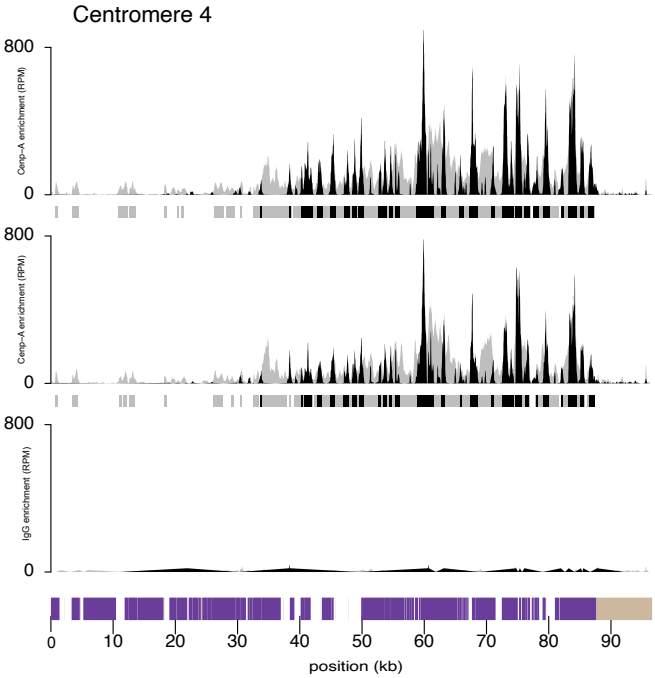
SupFig1

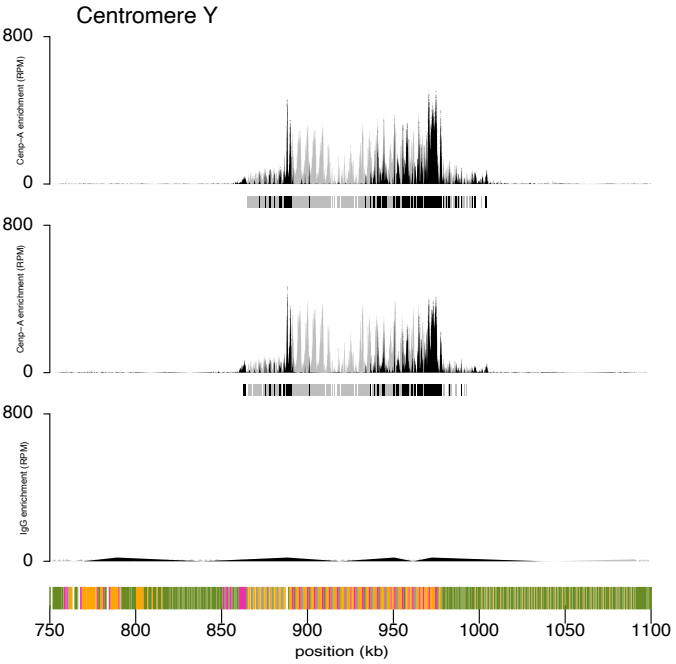
CUT&Tag results from the two additional CENP-A replicates (top two row) and the IgG negative control (bottom row) for each centromere in *D. simulans*. The y-axis represents the normalized CENP-A or IgG enrichment in RPM. Black and grey traits represent the enrichment based on uniquely and multi mapping reads respectively. The black and grey bars below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads respectively. The colored cytoband at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure.

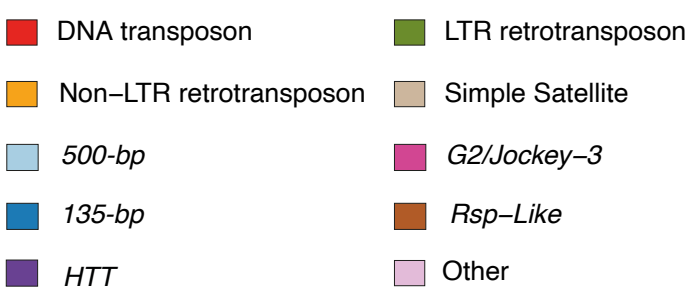

Centromere 2/3
Cenp−A enrichment (RPM)
IgG enrichment (RPM)
position (kb)


Centromere 2/3
Cenp−A enrichment (RPM)
IgG enrichment (RPM)
position (kb)


Centromere X
Cenp−A enrichment (RPM)
IgG enrichment (RPM)
position (kb)

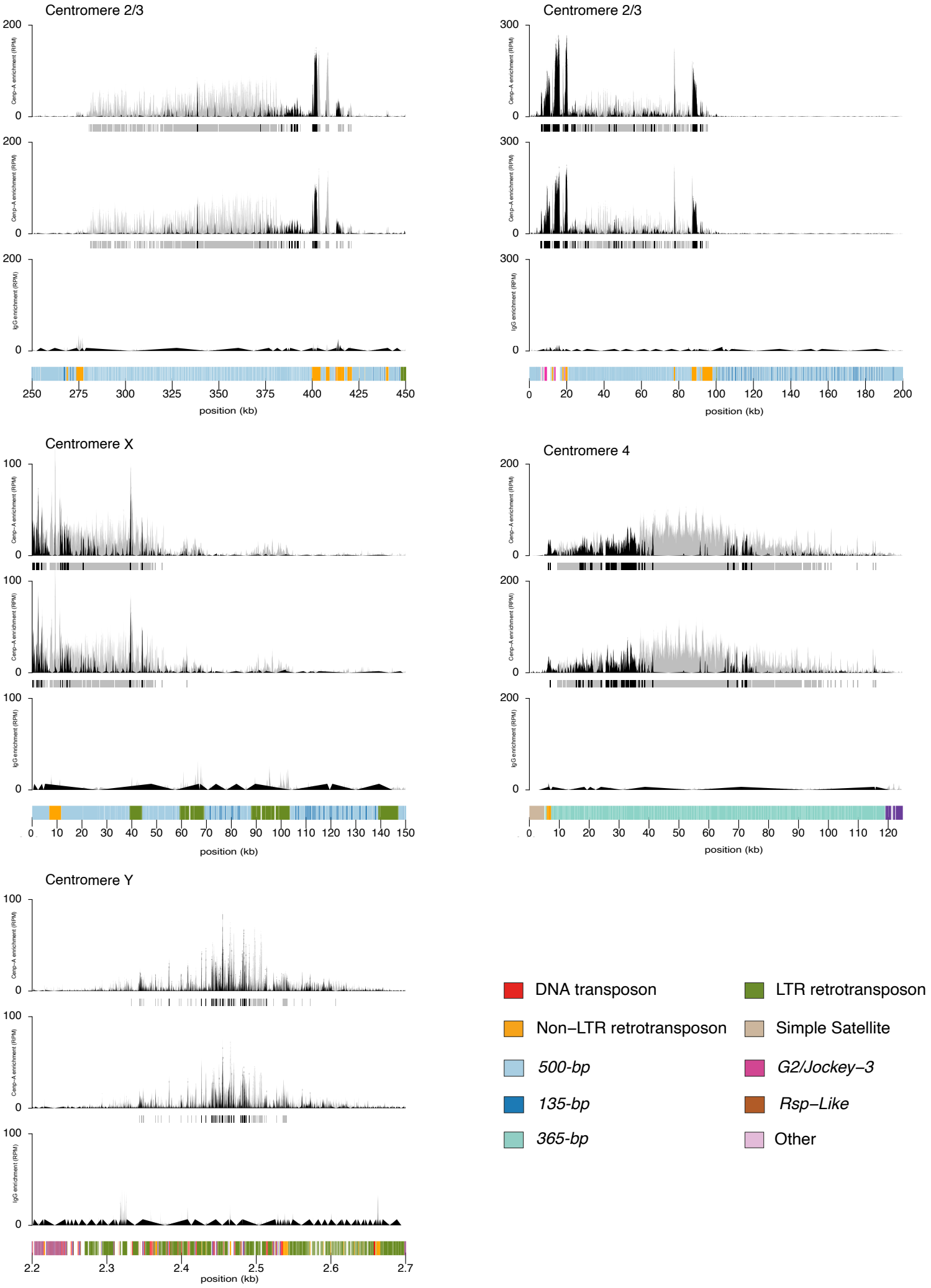

Centromere 4
Cenp−A enrichment (RPM)
IgG enrichment (RPM)
position (kb)


Centromere Y
Cenp−A enrichment (RPM)
IgG enrichment (RPM)
position (kb)


DNA transposon
LTR retrotransposon
Non−LTR retrotransposon
Simple Satellite
500-bp
G2/Jockey−3
135-bp
Rsp−Like
HTT
Other

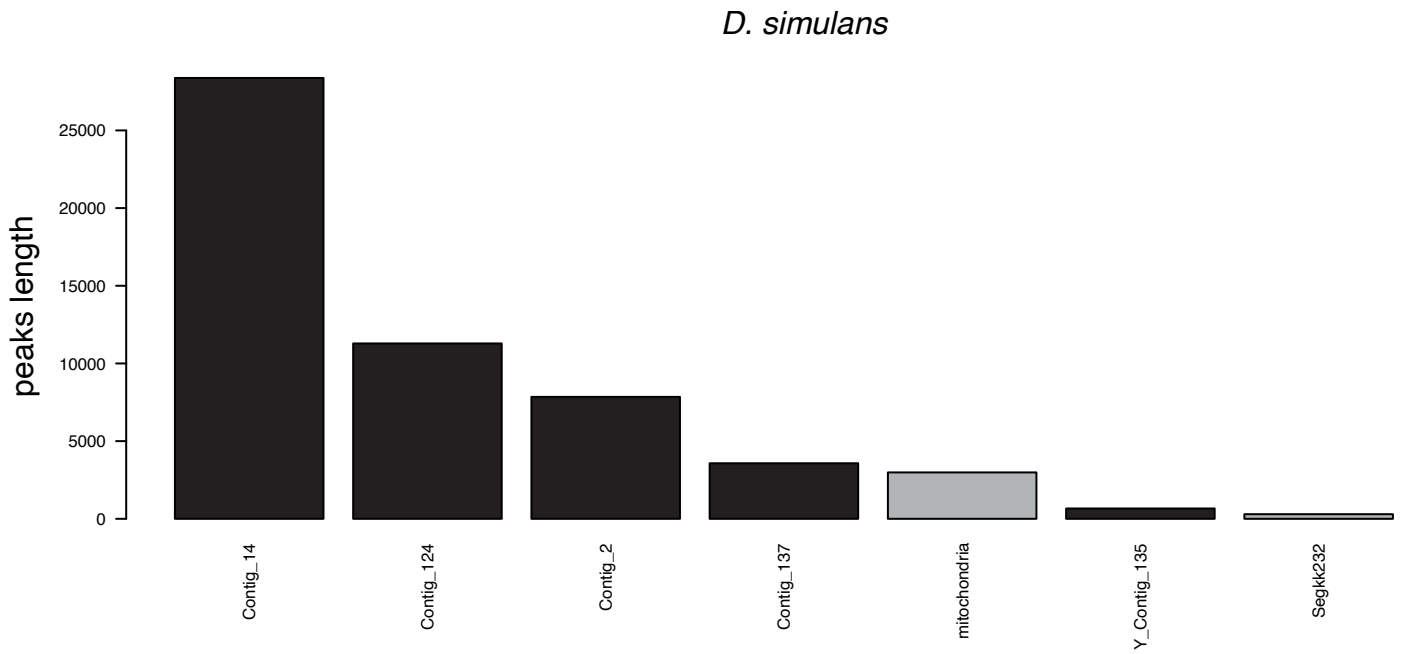
SupFig2

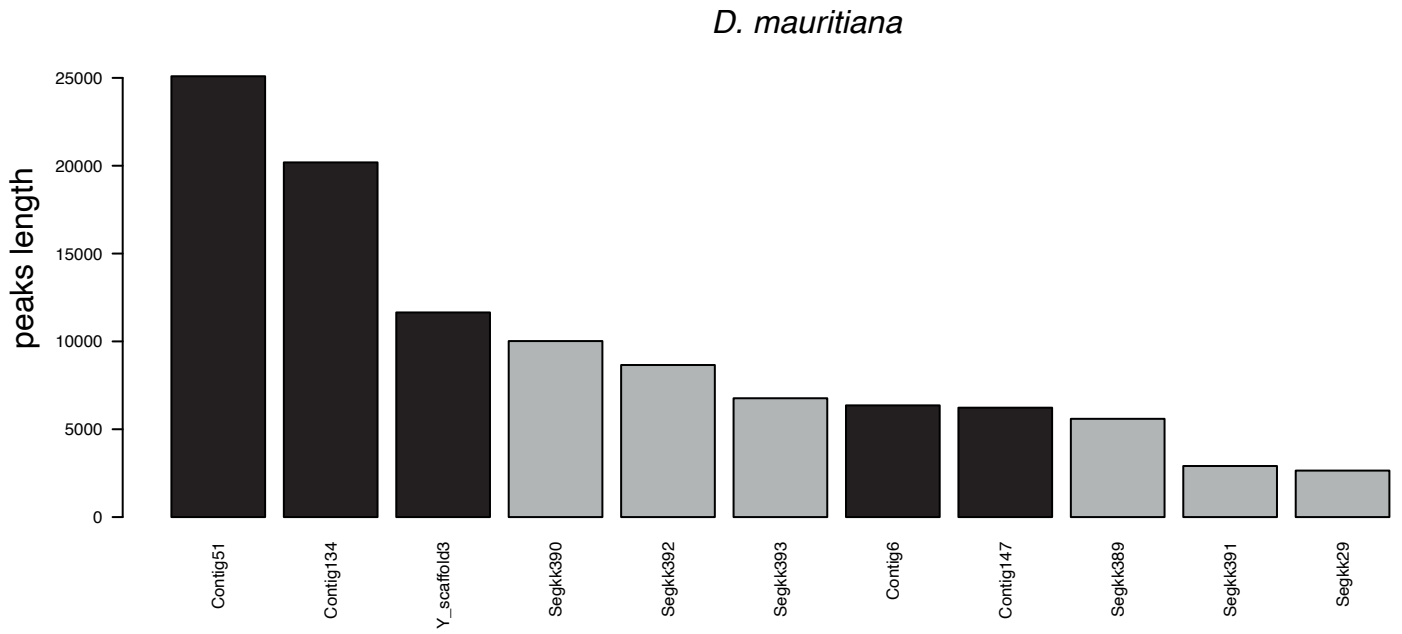
CUT&Tag results from the two additional CENP-A replicates (top two row) and the IgG negative control (bottom row) for each centromere in *D. sechellia*. The y-axis represents the normalized CENP-A or IgG enrichment in RPM. Black and gray traits represent the enrichment based on uniquely and multi mapping reads respectively. The black and grey bars below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads respectively. The colored cytoband at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure.

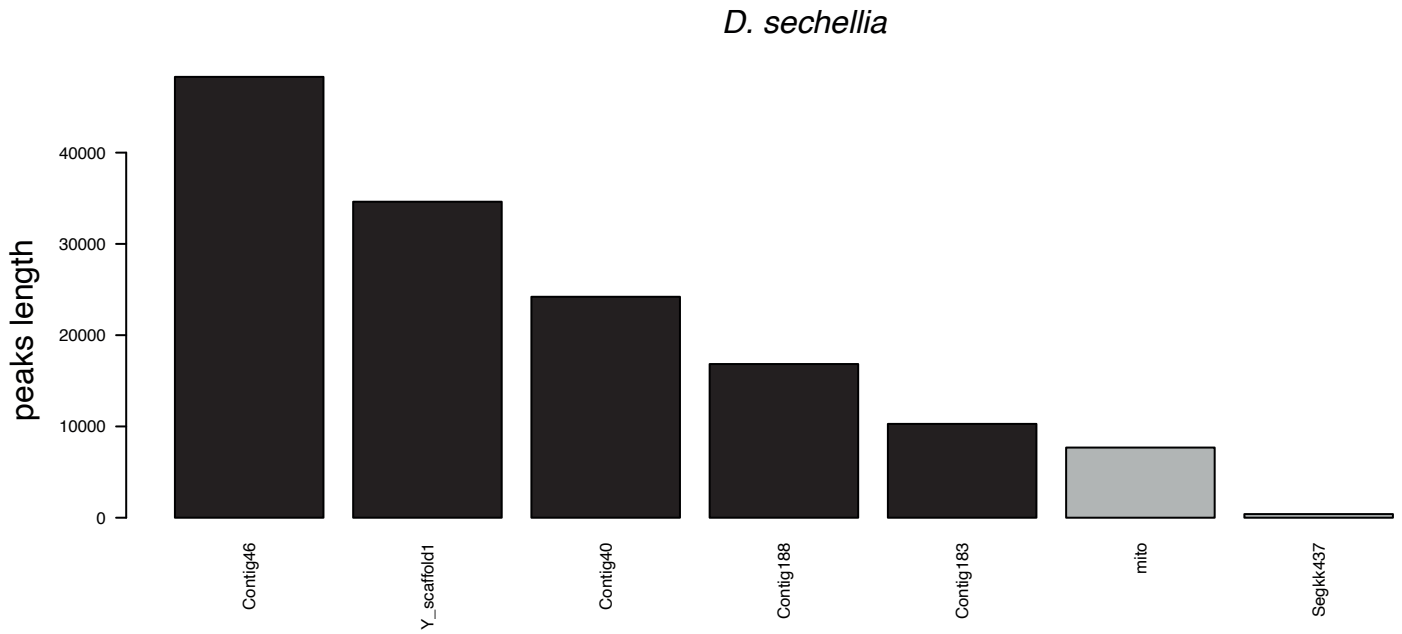

Centromere 2/3
Cenp–A enrichment (RPM)
IgG enrichment (RPM)
position (kb)
Centromere 2/3
Centromere X
Centromere 4
Centromere Y
DNA transposon
LTR retrotransposon
Non–LTR retrotransposon
Simple Satellite
500-bp
G2/Jockey–3
135-bp
Rsp–Like
365-bp
Other

SupFig3

CUT&Tag results from the two additional CENP-A replicates (top two row) and the IgG negative control (bottom row) for each centromere in *D. mauritiana*. The y-axis represents the normalized CENP-A or IgG enrichment in RPM. Black and grey traits represent the enrichment based on uniquely and multi mapping reads respectively. The black and grey bars below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads respectively. The colored cytoband at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure.

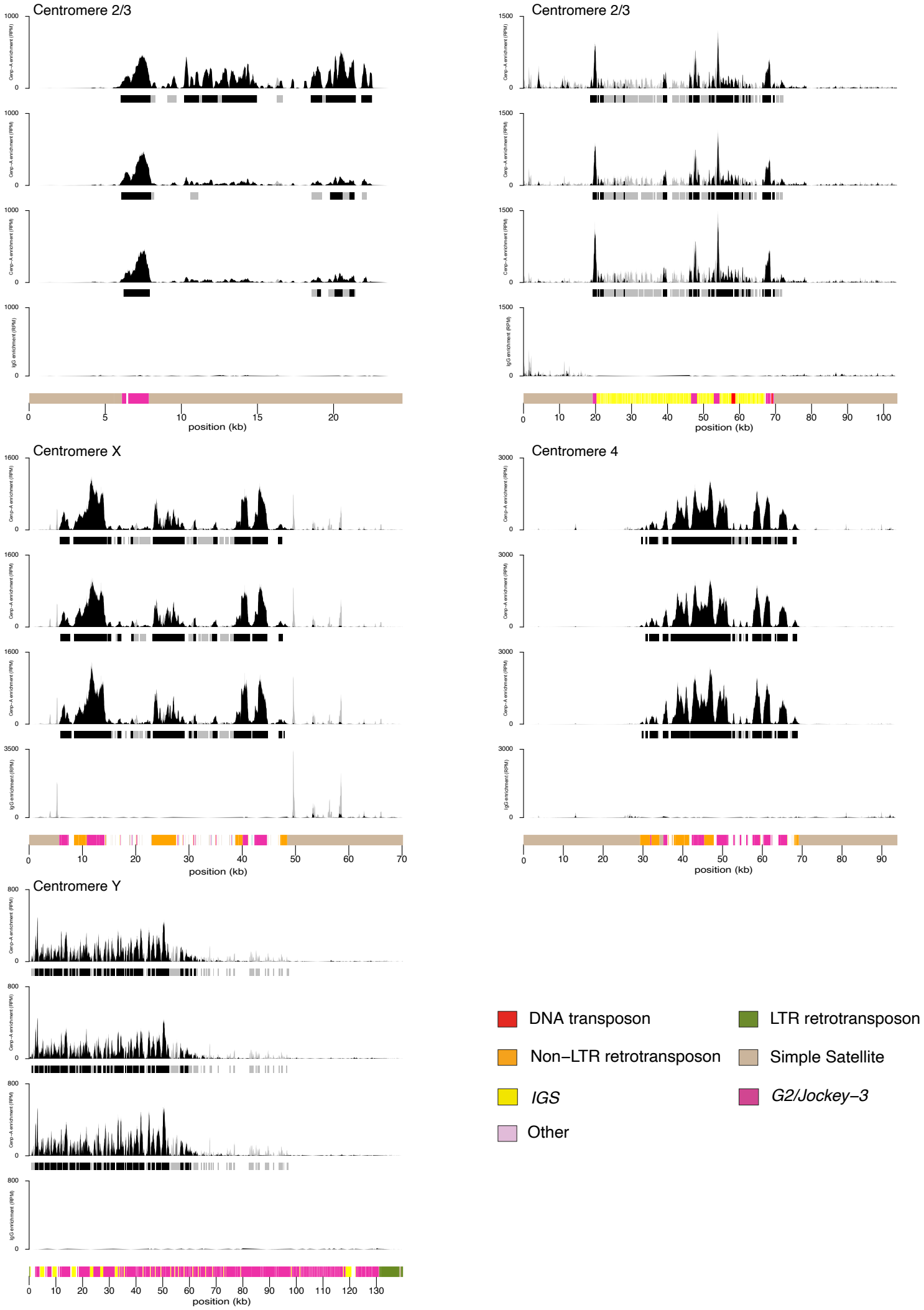

D. simulans
peaks length
25000
20000
15000
10000
5000
0
Contig_14
Contig_124
Contig_2
Contig_137
mitochondria
Y_Contig_135
Segkk232


D. mauritiana
peaks length
25000
20000
15000
10000
5000
0
Contig51
Contig134
Y_scaffold3
Segkk390
Segkk392
Segkk393
Contig6
Contig147
Segkk389
Segkk391
Segkk29


D. sechellia
peaks length
40000
30000
20000
10000
0
Contig46
Y_scaffold1
Contig40
Contig188
Contig183
mito
Segkk437

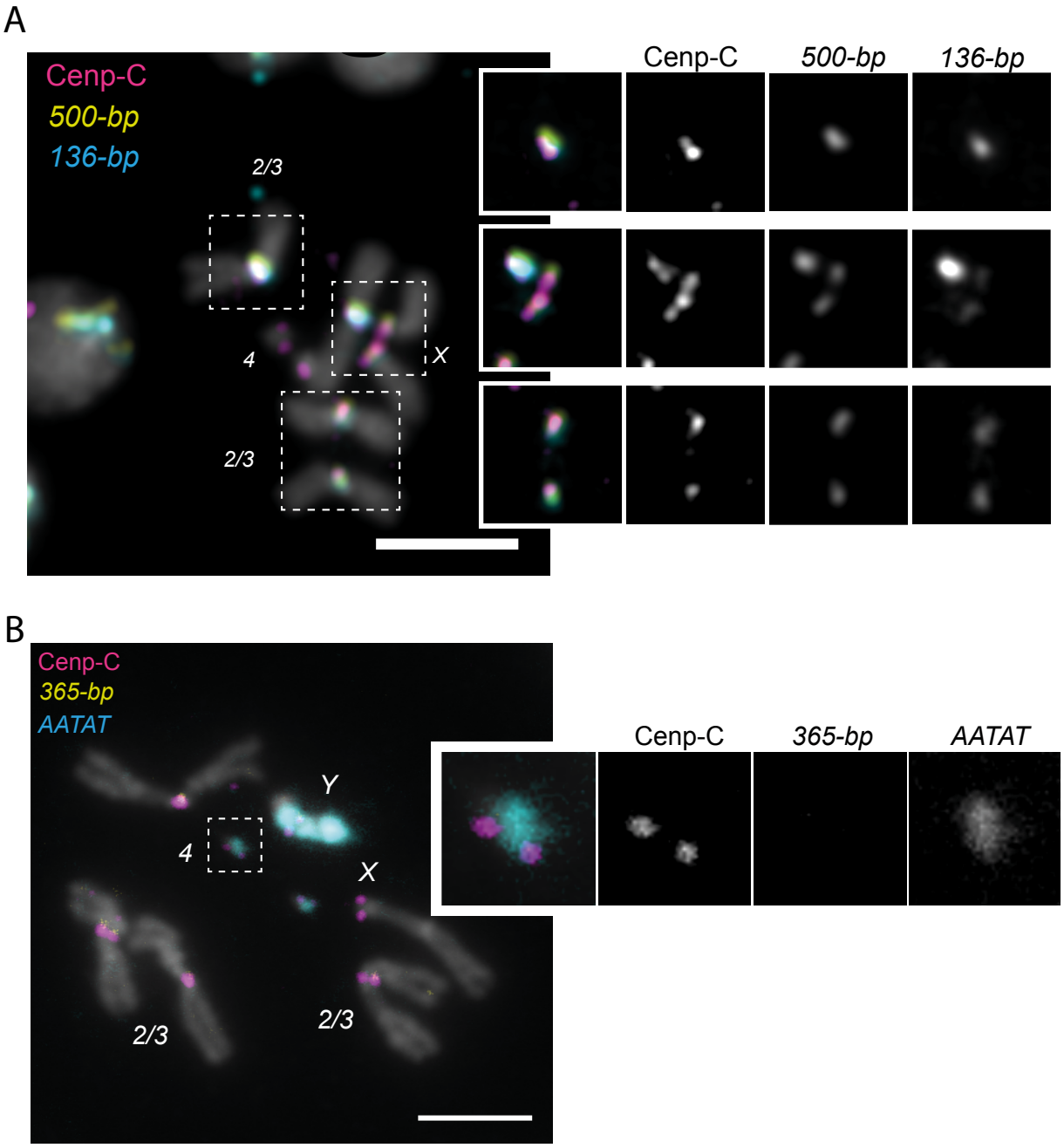
SupFig4

Location of the peaks resulting from the IDR analysis - significantly enriched region conserved between the three replicates. The y axis represents the sum of the peaks length for each contig. The contig corresponding to the centromere are colored in black.


Centromere 2/3
Centromere 2/3
Centromere X
Centromere 4
Centromere Y
Cenp-A enrichment (RPM)
IgG enrichment (RPM)
position (kb)
DNA transposon
LTR retrotransposon
Non–LTR retrotransposon
Simple Satellite
IGS
G2/Jockey–3
Other

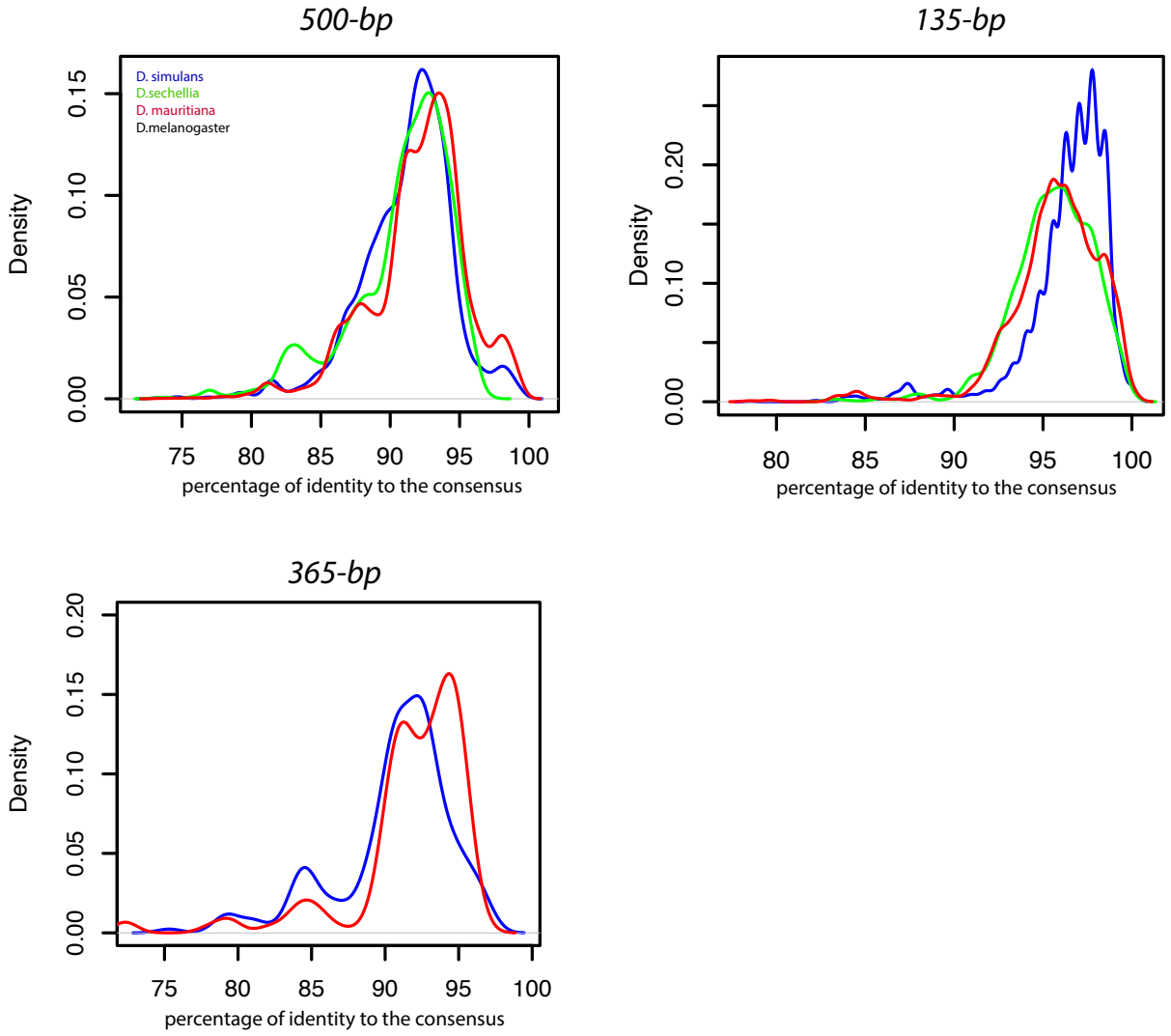
SupFig5

CUT&Tag results from the three CENP-A replicates (top two row) and the IgG negative control (bottom row) for each centromere in *D. melanogaster*. The y-axis represents the normalized CENP-A or IgG enrichment in RPM. Black and gray traits represent the enrichment based on uniquely and multi mapping reads respectively. The black and gray bars below each plot correspond to MACS2 peaks showing significantly enriched regions based on the unique and multi mapping reads respectively. The colored cytoband at the bottom of the plot shows the repeat organization. The color code is shown in the legend at the bottom of the figure.



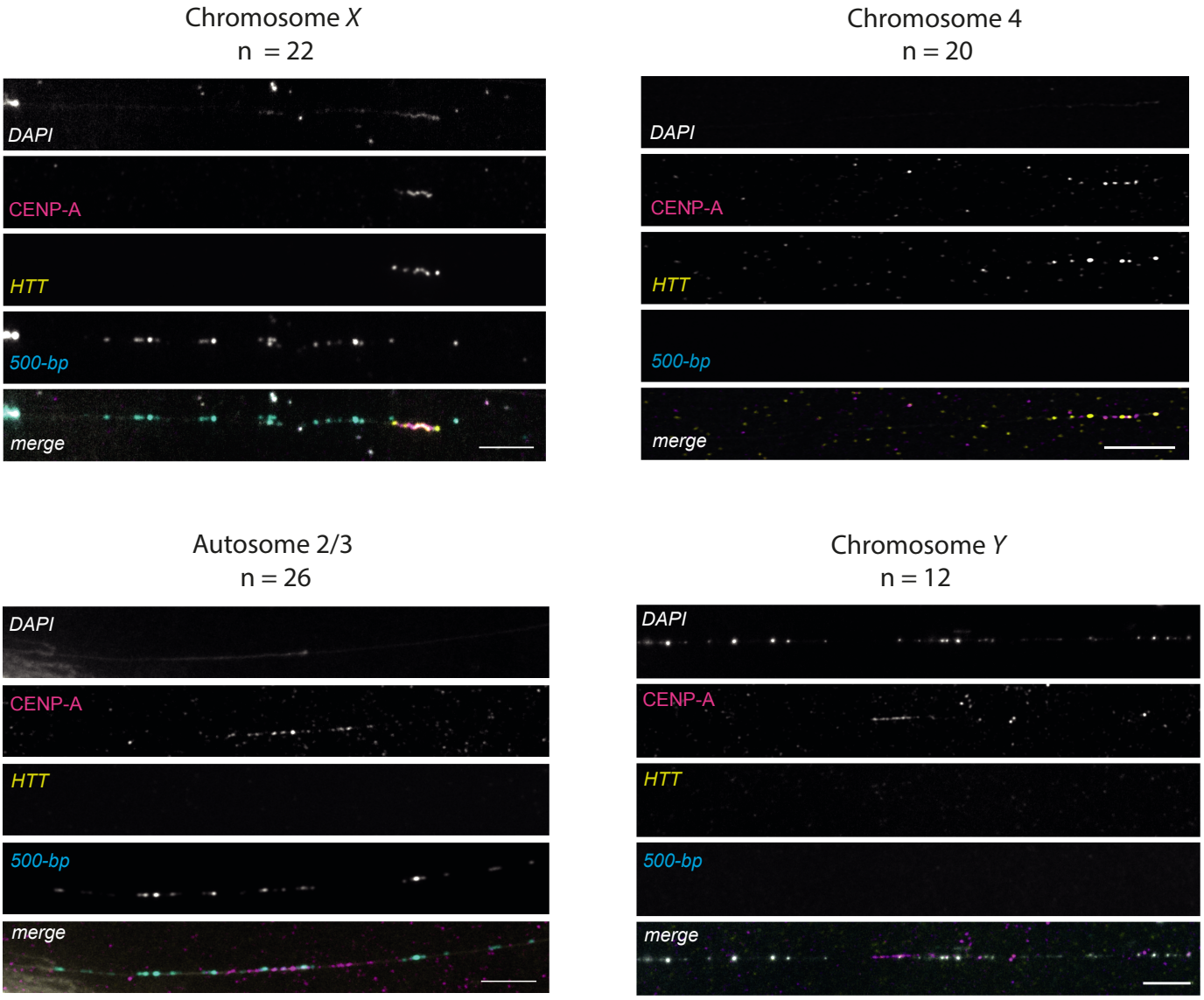
SupFig6

A.IF-FISH on mitotic chromosomes from the larval brain with CENP-C antibody and *500bp* and *136-bp* probes. The inset represents a zoom on each centromere. B. IF-FISH on mitotic chromosomes from the larval brain from *D. sechellia* with CENP-C antibody and *365-bp* and *AATAT* probes. The inset represents a zoom on the dot chromosome centromere.



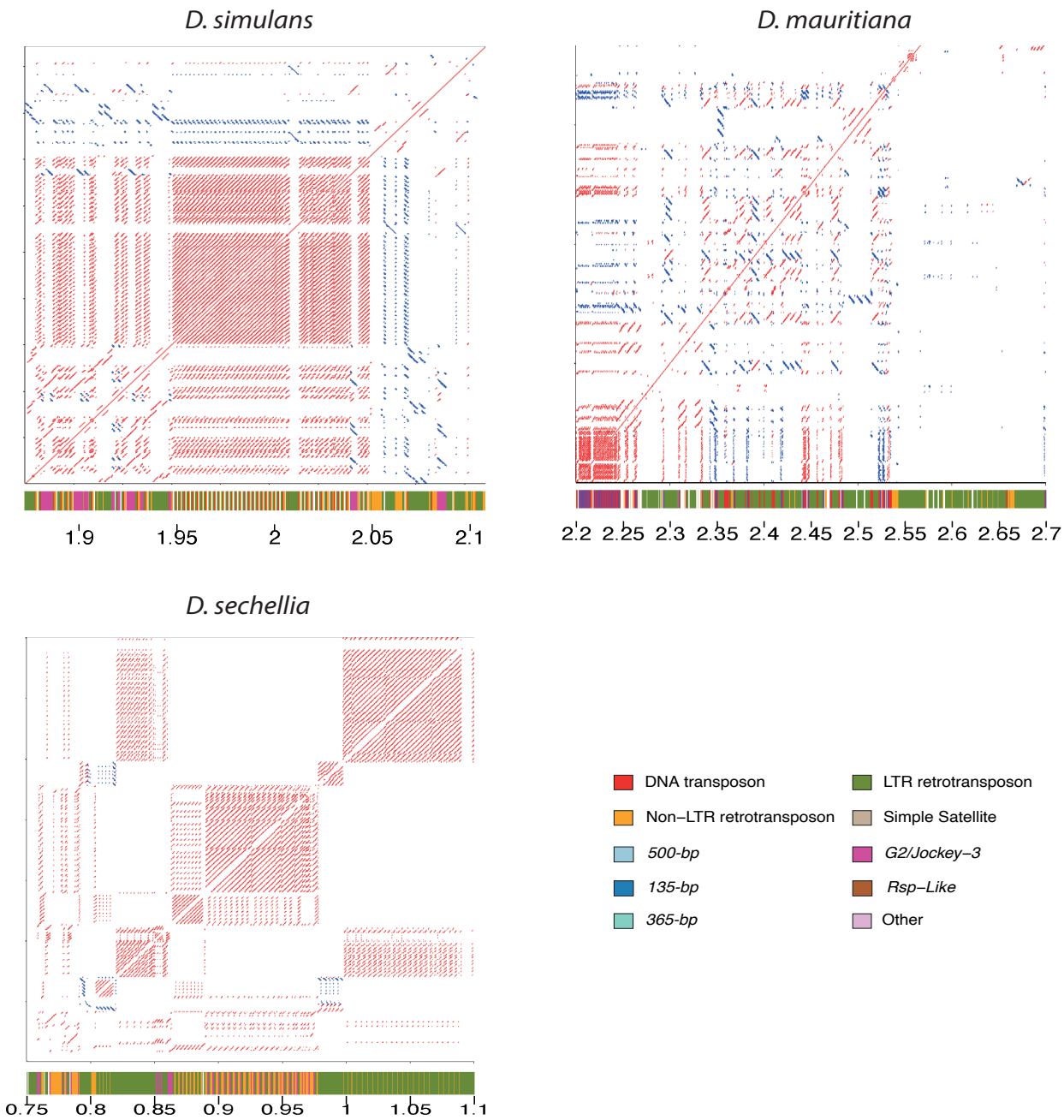
SupFig7

Distribution of the percentage of divergence of individual insertion from the consensus sequence for each centromeric satellite. Only insertions with a length > 80% of consensus length were kept. The percentage of divergence was extracted from the Blast output.



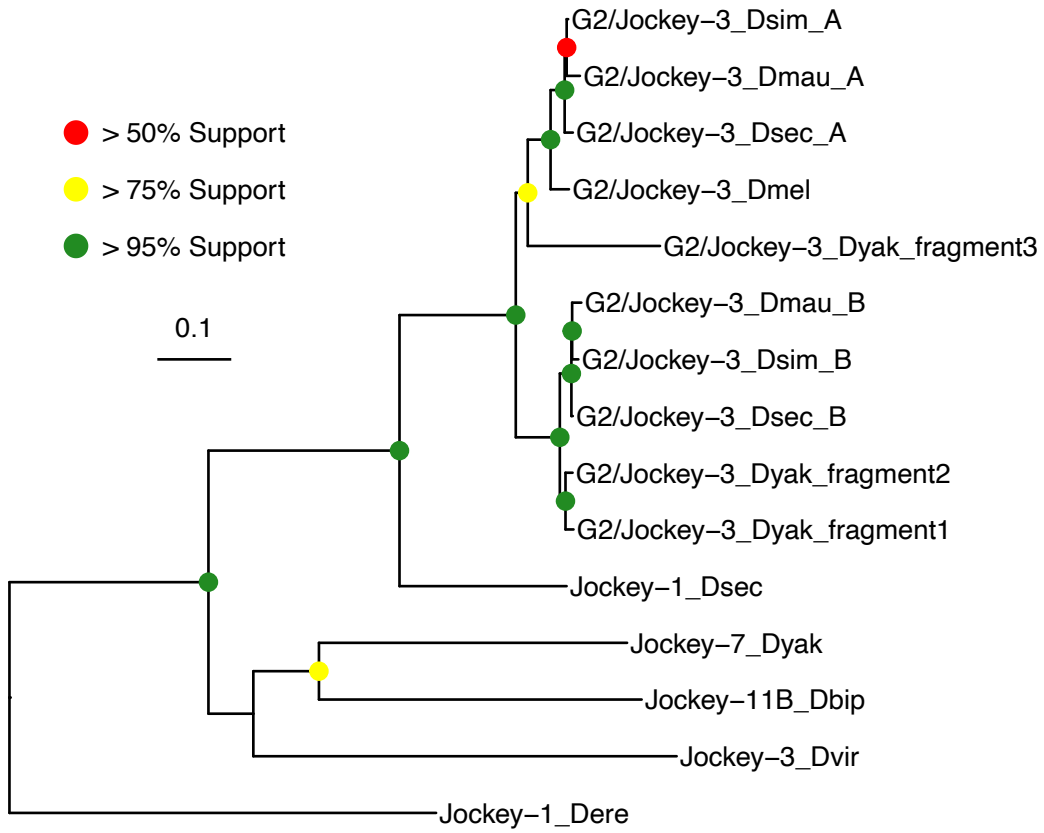
SupFig8

IF-FISH on chromatin fibers from the *D. sechellia* larval brains with CENP-A antibody and *500bp* and *HTT* probes. A representative image of each centromere pattern is presented along with the total number of images collected for each pattern. CENP-A is present on the HTT region with or without *500-bp* flanking, corresponding to the X and dot chromosome, respectively. CENP-A is also present on a *500bp* region, corresponding to the autosomal centromeres and without 500-bp nearby, consistent with the Y chromosome.



SupFig9

Dotplot from the alignment on the Y chromosome centromere against itself to highlight higher order repeat. The Dotplot were produced using re-DOT-able with a sliding window of 100bp. the cytoband below each dotplot represent the repeat composition of the region. The color code is indicated in the legend.



SupFig10

Phylogenetic tree with node support of consensus *G2/Jockey-3* ORF sequences in relation to closely-related Jockey elements. Closely-related Jockey elements were identified from[41]. Three *D. yakuba* fragments which span the >50% of the ORF are also included.