Dynamic turnover of centromeres drives karyotype evolution in Drosophila

Ryan Bracewell, Kamalakar Chatla, Matthew J. Nalley and Doris Bachtrog

Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA

Centromeres are the basic unit for chromosome inheritance, but their evolutionary dynamics is poorly understood. We generate high-quality reference genomes for multiple *Drosophila obscura* group species to reconstruct karyotype evolution. All chromosomes in this lineage were ancestrally telocentric and the creation of metacentric chromosomes in some species was driven by *de novo* seeding of new centromeres at ancestrally gene-rich regions, independently of chromosomal rearrangements. The emergence of centromeres resulted in a drastic size increase due to repeat accumulation, and dozens of genes previously located in euchromatin are now embedded in pericentromeric heterochromatin. Metacentric chromosomes secondarily became telocentric in the *pseudoobscura* subgroup through centromere repositioning and a pericentric inversion. The former (peri)centric sequences left behind shrunk dramatically in size after their inactivation, yet contain remnants of their evolutionary past, including increased repeat-content and heterochromatic environment. Centromere movements are accompanied by rapid turnover of the major satellite DNA detected in (peri)centromeric regions.

Introduction

Centromeres are the chromosomal regions to which the spindle microtubules attach during each cell division to ensure disjunction of chromosomes. In many eukaryotes, centromeres consist of species-specific kilobase- to megabase-sized arrays of tandem repeats (Willard 1990; Birchler *et al.* 2009; Wang *et al.* 2009; Melters *et al.* 2013). Centromeres are characterized by a distinct type of chromatin where histone H3 is replaced by a centromere-specific histone H3 variant (cenH3) (Karpen and Allshire 1997; Allshire and Karpen 2008), which interacts with other proteins to seed the kinetochore.

The specific chromosomal regions of centromere formation are typically highly repetitive and embedded in heterochromatin and are thus poorly studied in many species, and the mechanisms dictating centromere specification have remained elusive in most organisms. In budding yeast, characteristic sequences of ~120-bp fully specify centromere identity (Hieter *et al.* 1985). In organisms with complex repetitive centromeres, the site of kinetochore assembly is not strictly governed by primary DNA sequence but appears to be

epigenetically determined (Karpen and Allshire 1997; Allshire and Karpen 2008). The key-determining factor in epigenetic centromere specification is the deposition of the centromere-specific histone H3 variant cenH3. Numerous factors have been identified that are required for cenH3 localization, but the mechanisms of centromere specification are not known in most species. Non-coding transcripts are associated with centromeres, and transcription might be linked to cenH3 chromatin assembly in some systems (Chen *et al.* 2015; McNulty *et al.* 2017; Talbert and Henikoff 2018). The concept of epigenetic specification of the centromere is further supported by the existence of neo-centromeres, which form at ectopic chromosomal loci in many species (Sart *et al.* 1997) and lack shared sequence features (Burrack and Berman 2012). Several studies have found that centromeric satellite DNA may form unusual secondary structures (termed non-B-form secondary structures, as opposed to the canonical B-form structure first described by Watson and Crick), and various types of non-B-form structures have been observed both *in vitro* and *in vivo* in centromeric DNA of various organisms, including single-stranded DNA, hairpins, R-loops and i-motifs (Garavís *et al.* 2015; Kabeche *et al.* 2018; Kasinathan and Henikoff 2018). Thus, multiple factors including epigenetic marks, primary DNA sequence and secondary structure may act in concert to ensure the faithful formation of a centromere on each chromosome.

Centromeres are the basic unit for chromosome inheritance, but their evolution is highly dynamic. Both the centromeric histone protein (cenH3, which is called *Cid* in Drosophila), as well as the underlying DNA sequence that forms the centromere (that is, the centromeric satellite DNA) vary among species (Henikoff *et al.* 2001). Centromere repeats have been proposed to act as selfish genetic elements by driving non-Mendelian chromosome transmission during meiosis, thus prodding the rapid evolution of centromeric proteins to restore fair segregation (Henikoff *et al.* 2001; Malik and Henikoff 2009).

In addition to their primary sequence, the position of centromeres along the chromosome varies widely as well and results in diverse karyotypes across species (Tree of Sex Consortium 2014). Centromeres can be found at the end of chromosomes or in the center (telocentric versus metacentric chromosomes). Changes in centromere position can be driven by chromosomal re-arrangements, moving an existing centromere to a new location, or by evolutionary centromere repositioning – a move of the centromere along the chromosome not accompanied by structural rearrangements (Ferreri *et al.* 2005; Carbone *et al.* 2006; Rocchi *et al.* 2012; Schubert 2018; Nishimura *et al.* 2019). Thus, either pericentric inversions, or neo-centromere formation, i.e. the seeding of a novel centromere in a region previously not containing a centromere function, can drive karyotype evolution.

Studies of repetitive DNA, and the centromere in particular, are notoriously difficult. Most of our knowledge on centromere evolution is based on cytological studies, and comparison of chromosome morphology and mapping of repetitive DNA has provided a rich picture of the diversity of karyotypes. Karyotypes have been extensively investigated in Drosophila species, and the gene content of chromosome arms is conserved in flies (termed Muller elements; Muller 1940; Patterson and Stone 1952; Ashburner 1989; Whiting *et al.* 1989). The ancestral karyotype of Drosophila consists of five large telocentric chromosomes (rods; termed Muller elements A-E) and the much smaller dot chromosome (Muller F), and novel karyotypes have originated repeatedly by chromosomal fusions and inversions. Flies in the *obscura* group harbor a diversity of karyotypes (**Fig. 1;** Sturtevant 1936; Patterson and Stone 1952; Buzzati-Traverso and Scossiroli 1955; Powell 1997). The ancestral karyotype (found in flies of the *subobscura* group) consists of five ancestral telocentric rods and the dot chromosome, while species in the *obscura* and *affinis* group mainly contain metacentric chromosomes, and most chromosomes in the *pseudoobscura* group are telocentric. Both pericentric inversions and chromosomal fusions were invoked to explain this diversity in karyotypes (Patterson and Stone 1952; Buzzati-Traverso and Scossiroli 1955; Segarra *et al.* 1995; Schaeffer *et al.* 2008). Here, we use long-read sequencing techniques to assemble high-quality genomes for multiple species in the *obscura* group with different karyotypes, to reconstruct the evolutionary history of karyotypic evolution at the molecular level. Our assemblies recover entire chromosomes, including large fractions of repetitive DNA and pericentromere sequences and centromere-associated repeats. We uncover a dynamic history of centromere evolution in this species group, and identify centromeres likely being formed *de novo* in gene-rich regions, followed by dramatic size increases due to accumulation of repetitive DNA. In some species, these novel centromeres become obsolete secondarily, and centromere inactivation is accompanied by loss of repetitive DNA while maintaining a heterochromatic configuration. The transitions in karyotypes are associated with rapid turnover of centromere-associated repeats between species and suggest an important role for repetitive DNA in the evolution of reproductive isolation and the formation of new species.

Results

Long-read sequencing allows assembly of entire chromosomes, including pericentromeric regions

Centromeres in many multicellular eukaryotes consist of highly repetitive satellite DNA, and are typically embedded in repeat-rich heterochromatin (that is, the pericentromeric heterochromatin; Willard 1990; Birchler *et al.* 2009; Wang *et al.* 2009; Melters *et al.* 2013). Investigation of centromere biology has been

100 greatly hampered by a lack of high-quality genome sequences (Talbert *et al.* 2018; Chang *et al.* 2019). Here,
101 we obtain high-quality sequence assemblies of several Drosophila species in the *obscura* group, to
102 reconstruct karyotype evolution at the molecular level.
103
104 We used a combination of PacBio and Nanopore sequencing and Hi-C scaffolding to generate highly
105 contiguous genome assemblies of high quality for *D. pseudoobscura*, *D. athabasca*, *D. subobscura* and *D.*
106 *lowei* (**Supplementary File 1, Supplementary File 2**, see Methods and **Appendix 1**). We also included our
107 recently published high-quality assembly of *D. miranda* to this analysis (Mahajan *et al.* 2018). For each
108 species, we were able to generate chromosome-level assemblies that contain large stretches of repetitive
109 DNA, including pericentromere and putative centromere sequence (**Fig. 2**, **Figure 2 - figure supplement 1,**
110 **Figure 2 - figure supplement 2**). In particular, euchromatic regions from each chromosome were typically
111 assembled into just a few contigs (**Table 1, Fig. 2**), and Hi-C scaffolding allowed for the orientation and
112 placement of these contigs into chromosome-level assemblies (**Figure 2 - figure supplement 3**). BUSCO
113 (Benchmarking Universal Single-Copy Orthologs) scores reveal that our genomes are highly complete
114 (**Supplementary File 3**) and match or exceed other recently published high quality Drosophila genomes
115 (Mahajan *et al.* 2018; Puerma *et al.* 2018; Miller *et al.* 2018; Karageorgiou *et al.* 2019). Identification of *D.*
116 *melanogaster* orthologs in our genome assemblies allowed us to unambiguously assign each chromosome
117 to a Muller element (**Fig. 3A**) and further confirm the high level of gene content conservation across
118 chromosomes in Drosophila (Muller 1940; Patterson and Stone 1952; Bhutkar *et al.* 2008). Our
119 chromosome-level genome assemblies are in agreement with the karyotypes based on cytology (Buzzati-
120 Traverso and Scossiroli 1955; see **Figs. 1-4**). Chromatin interaction maps display high levels of associations
121 within chromosome arms and lower levels between (Lieberman-Aiden *et al.* 2009), and our Hi-C scaffolding
122 maps reconstruct five large telocentric chromosomes in *D. subobscura*, three large metacentric and two
123 telocentric chromosomes in *D. athabasca*, two large metacentric and three telocentric chromosomes in *D.*
124 *lowei* and four large telocentric and one metacentric chromosome in *D. pseudoobscura* and *D. miranda*
125 (**Figure 2 - figure supplement 3**).
126
127 We used transcriptome data for several species to aid in genome annotation (**Supplementary File 1**,
128 **Supplementary File 4**). Our annotations identified 12,714 protein-coding genes in *D. subobscura*, 13,665 in
129 *D. athabasca*, 14,547 in *D. lowei*, and 14,334 in *D. pseudoobscura*. The number of annotated genes in *D.*
130 *subobscura* is very similar to the 13,317 protein-coding genes in another assembly of *D. subobscura*
131 (Karageorgiou *et al.* 2019) and the 13,453 genes of its close relative *D. guanche* (Puerma *et al.* 2018).
132 Further, the number of annotated genes in *D. pseudoobscura* and *D. lowei* (**Supplementary File 4**) are

similar to the number in the current version of *D. pseudoobscura* (14,574 genes in Dpse_3.0). Currently, no high-quality genome assembly or annotation exists for a species that is closely related to *D. athabasca* (*affinis* subgroup), but our *D. athabasca* annotation is largely in line with the other species from the *obscura* group. We generated *de novo* repeat libraries for each species and concatenated them with a published *Drosophila* repeat library (Bao *et al.* 2015) to mask repeats in each genome. Our assemblies contained large amounts of repetitive DNA; we repeat-masked a total of 10.2 Mb, 54.4 Mb, 46.5 Mb, and 28.3 Mb from the assembled chromosomes of *D. subobscura*, *D. athabasca*, *D. lowei* and *D. pseudoobscura*, respectively (**Supplementary File 5**).

As expected, gene and repeat density are variable across all chromosomes in all species and show opposite patterns (**Fig. 2**). We see sharp decreases in gene density near one end at telocentric chromosomes and near the center of metacentric chromosomes, while repeat density peaks in gene poor regions (**Fig. 2**). Thus, our assemblies contain substantial amounts of (peri)centromeric repetitive DNA. We used repeat-enrichment to define the boundaries of euchromatic chromosome arms versus pericentromeric repeats for each species using a cut-off of 20% (*D. subobscura*) or 40% (*D. athabasca*, *D. pseudoobscura*, *D. miranda* and *D. lowei*) repeat-masked DNA across 100kb sliding windows away from the centromere (**Fig. 2**). A lower cut-off was used for *D. subobscura*, as this species has a very small, compact and repeat-poor genome (see also Karageorgiou *et al.* 2019).

The sizes of (peri)centromere regions recovered for each chromosome and species vary considerably (**Table 1**). In most species, the identified pericentromeric regions are almost entirely consisting of masked repetitive DNA (**Fig. 2**, **Supplementary File 5**), but pericentromeric regions in the *D. subobscura* assembly tend to be less repetitive (mean = 45.5%, **Fig. 2**). The transition from repetitive pericentromere to euchromatic chromosome arms is also much less abrupt in this species (**Fig. 2**). Genome-wide alignments between our *D. subobscura* assembly and a recently published high-quality genome confirm our inference of pericentromeric DNA (Puerma *et al.* 2018; Karageorgiou *et al.* 2019) and verify the orientation of the chromosomes (**Figure 3 - figure supplement 1**). Further, comparison with another genome assembly in the *obscura* group (*D. guanche*; **Figure 3 - figure supplement 1**) verifies known large chromosomal inversions between species and/or strains (Orengo *et al.* 2019) and provides further evidence of the quality of our assemblies.

We generated genome-wide H3K9me3 profiles for several species, which is a histone modification typical of heterochromatin (Elgin and Reuter 2013). Indeed, we see significant enrichment of heterochromatin (as

measured by H3K9me3) at pericentric regions (**Fig. 2**). Thus, combined patterns of gene density, repeat content, and heterochromatin enrichment together with comparisons to published genomes confirm that our high-quality genome assemblies have allowed us to assemble large stretches of the pericentromere and putative centromeric regions in the various Drosophila species.

Karyotype evolution in the *obscura* species group

The evolutionary relationships in this species group suggest that the ancestor of the *obscura* species group had the basic Drosophila karyotype consisting of five large telocentric rods and the dot chromosome (Patterson and Stone 1952; Buzzati-Traverso and Scossiroli 1955; Gao *et al.* 2007). Flies in the *subobscura* subgroup (which includes *Drosophila subobscura*, and its close relative *D. guanche*) harbor this ancestral chromosome configuration (Patterson and Stone 1952; Powell 1997), and whole genome alignments between these two species show long tracks of collinearity (**Figure 3 - figure supplement 1**). In species of the *obscura* subgroup, Muller element A (which is the X chromosome in Drosophila), B and E have become metacentric (and in several members of this subgroup, element C and D have been fused; Gao *et al.* 1949). In an ancestor of the *pseudoobscura* and *affinis* subgroups, Muller element A and D fused. Previous research suggested that the fusion between Muller A and D involved two telocentric chromosomes, followed by a pericentric inversion translocating genetic material from Muller A to Muller D (Segarra and Aguadé 1992; Segarra *et al.* 1995; note that we refer to these Muller elements as Muller A-AD). Phylogenetic considerations and genomic analysis, however, make a scenario of a fusion between an telocentric element D and a metacentric Muller element A more likely (see below; and also Schaeffer *et al.* 2008). Species in the *affinis* subgroup, which include *D. athabasca* and *D. affinis*, have maintained this karyotype where Muller A-AD are fused, and Muller B and E are metacentric. In some species of the *pseudoobscura* subgroup (that is, *D. miranda*, *D. pseudoobscura* and *D. persimilis*), however, both Muller B and E are telocentric, and parsimonious reconstruction suggests that these telocentric chromosomes evolved from a metacentric ancestor. Karyotyping in the little investigated species *D. lowei*, which is an outgroup to *D. miranda*, *D. pseudoobscura* and *D. persimilis*, suggests that only one of these autosomes is telocentric, while the other is still metacentric (Heed *et al.* 1969). Thus, the *obscura* group of Drosophila provides an exciting opportunity to study the evolution of vastly different karyotypes: All chromosomes in this group were ancestrally telocentric, three Muller elements became metacentric and two of these metacentric chromosomes secondarily reverted back to telocentric chromosomes.

Our high-quality genome assemblies recapitulate these large-scale changes in chromosome morphology, and comparison of gene content with *D. melanogaster* chromosomes allows us to unambiguously establish

chromosome arm homologies (**Fig. 3A**). Our assembly recovers five telocentric rods in *D. subobscura*, each flanked by repeat-rich pericentromeric DNA on one end. *D. athabasca*, in contrast, has three metacentric chromosomes, and our chromosome-level assembly recovers the (peri)centromeric regions in the center of these chromosomes (see **Fig. 2**). We assemble ~14 Mb of pericentric DNA on Muller A-AD, ~23 Mb on Muller B, and ~11 Mb on Muller E. In *D. lowei*, we assemble two telocentric large rods, and two metacentric chromosomes (Muller A-AD and E), and our assembly contains about 17 Mb of pericentric DNA on Muller A-AD and 15 Mb on Muller E (**Table 1**). *D. miranda* and *D. pseudoobscura*, in contrast, contain only a single metacentric chromosome (Muller A-AD) with 21 Mb and 14 Mb of assembled pericentromere sequence.

Identification of centromere-associated repeats

Centromeres of many eukaryotes are associated with large stretches of satellite DNA, but simple satellites are typically missing from most assemblies. However, long-read sequencing technology has allowed resolving the detailed structure of complex satellites in *D. melanogaster* (Khost *et al.* 2017; Chang *et al.* 2019), and probing of high quality whole-genome assemblies based on long-read data has allowed for the recovery of putative centromere satellite motifs in *D. miranda* (Mahajan *et al.* 2018). In addition, given the extremely high repeat abundance of centromere repeats (often >10,000 copies per chromosome), the most abundant tandem repeat in a given genome is a prime candidate for the centromere repeat (Melters *et al.* 2013), and bioinformatic methods to identify high-copy tandem repeats from both raw sequencing reads or fragmented assemblies have been applied to hundreds of species to study centromere evolution across animals and plants (Melters *et al.* 2013).

Here, we combine multiple complementary strategies to identify putative centromere-associated satellites in our Drosophila species (**Table 2**). We first used Tandem Repeat Finder (TRF; Benson 1999) to identify satellite sequences in our assemblies that are enriched in putative centromeric regions (i.e., ends of telocentric chromosomes and scaffolding breakpoints in the middle of metacentric chromosomes). Additionally, we identify short simple satellites (<20 bp) directly from raw Illumina reads using k-Seek (Wei *et al.* 2018), and we also identify tandem repeats directly from long-read data (Nanopore and PacBio) using TideHunter (Gao *et al.* 2019). The latter two approaches provided analyses of satellite and repeat enrichment independent of our genome assemblies.

We previously annotated satellites in the *D. miranda* genome and recovered two repeats, a 21-bp motif and a 99-bp motif that were the most abundant tandem repeats (Mahajan *et al.* 2018). Both of these satellites were enriched at the putative centromere of each chromosome in our assembly and had

sequence characteristics of typical centromere satellites (Mahajan *et al.* 2018). To identify putative centromere-associated repeats, we annotated satellites in our high-quality genomes and plotted the abundance of different length unique motifs (**Fig. 4A, B, Figure 4 - figure supplement 1**). Using this pipeline, our TRF analysis recovered the same 21-bp and 99-bp motif (or higher order variants of the two motifs) in *D. miranda*, with a higher order variant containing four 21-bp motifs (an 84-bp repeat) being the most common repeat unit (**Fig. 4A**). We identify the same 21-bp motif in *D. pseudoobscura* (again with an 84-mer being the most common higher order variant, **Fig. 4A**). Intriguingly, however, we fail to detect the 99-bp motif in *D. pseudoobscura*, both when probing for abundant satellites, and when searching the genome directly for the 99-mer. Similarly, we find a higher-order variant of the 21-mer in *D. lowei* (a 63-mer; **Fig. 4A**), and only a distantly related 99-mer on Muller A-AD (mean = 18.1% sequence divergence). The 21-mer is enriched in the (peri)centric regions of most chromosomes in *D. pseudoobscura* and *D. lowei* (**Fig. 4B, Figure 4 - figure supplement 1**), except for Muller A-AD and the dot chromosomes in *D. pseudoobscura*. K-Seek analyses aimed at identifying abundant simple satellites failed to identify any highly enriched candidates (**Table 2**, **Figure 4 - figure supplement 2**), consistent with previous work in *D. pseudoobscura* (Wei *et al.* 2018). Probing of PacBio and Nanopore long-reads with TideHunter was consistent with our TRF analyses and identified the 21-mer and higher-order variants in each of the three species (**Figure 4 - figure supplement 3**). Further, we confirm the 99-mer in *D. miranda* and its absence in *D. pseudoobscura* or *D. lowei* (**Table 2**, **Fig. S7**).

We used FISH hybridization to confirm the location of these satellites in the genomes of *D. miranda* and *D. pseudoobscura* (**Fig. 4C, Supplementary File 6, Figure 4 - figure supplement 4**). Indeed, we find both the 21-mer and the 99-mer enriched around the centromeres in *D. miranda* (**Fig. 4C**). In contrast, FISH with the same probes in *D. pseudoobscura* only showed strong staining for the 21-mer around its centromeres but we failed to detect any signal for the 99-mer (**Fig. 4C**). Consistent with our genome assembly, we failed to detect the 21-mer on Muller A-AD and the dot in *D. pseudoobscura* using FISH (**Fig. 4C**). Thus, FISH experiments confirm our bioinformatics analysis, and strongly suggests that the composition of the (peri)centromeric repeat landscape has changed between *D. miranda* and *D. pseudoobscura*. In particular, we infer that the 21-bp repeat was present in an ancestor of the *pseudoobscura* subgroup with a (peri)centric distribution, and that the 99-bp motif associated with (peri)centromeres in *D. miranda* was newly gained after their split.

In the *D. athabasca* genome, we identify a 160-bp motif as the most abundant repeat unit (**Fig. 4A**), and this motif is enriched in the assembled (peri)centromere in both telocentric and metacentric chromosomes

(**Fig. 4B, Figure 4 - figure supplement 1**). K-Seek and TideHunter analyses also identified a 11-mer (**Table 2**, **Figure 4 - figure supplement 2, Figure 4 - figure supplement 3**), which was less abundant in the genome assembly and in the raw PacBio reads than the 160-bp repeat (**Fig. 4A**, **Figure 4 - figure supplement 1, Figure 4 - figure supplement 3**). FISH experiments again confirm our bioinformatics approach and demonstrate that the 160-bp satellite is localized to the centromeric region on all chromosomes except Muller F (**Fig. 4C**). Lower intensity of fluorescence on Muller C suggests less satellite sequence on this chromosome (**Figure 4 - figure supplement 4**).

In the *D. suboobscura* assembly, on the other hand, we fail to detect any strong enrichment of a particular satellite using TRF (**Fig. 4A**). In its close relative *D. guanche*, in contrast, we clearly identify a known 290-bp centromere-associated satellite (Bachmann *et al.* 2009; Puerma *et al.* 2018). Thus, while *D. subobscura* has the most contiguous genome assembly of all species investigated (**Fig. 2**), this suggests that we did not assemble its centromere-associated repeat. Computational identification and quantification of short tandemly repeating sequences using k-Seek revealed a 12-bp satellite that is highly enriched in raw Illumina whole genome sequencing reads of *D. subobscura* and yet appears largely absent from our assembly (or from *D. guanche* and all other species; **Figure 4 - figure supplement 2**). Investigation of our Nanopore long reads also identified the 12-bp satellite and uncovered an additional 107-bp satellite in *D. subobscura* (**Table 2**, **Figure 4 - figure supplement 3**). Inspection of unplaced scaffolds in *D. subobscura* reveals one short contig (14.8 kb in size) with extremely high Illumina and Nanopore read coverage (650x and 8,500x in libraries sequenced to a genome wide average of 118x and 40x; **Figure 4 - figure supplement 5**), and this contig contains two large arrays of the 12-bp satellite (one containing 278 copies and the other 250 copies). In addition, we identify this 12-bp array at the start of our assembly of Muller B (**Fig. 4B**). Likewise, we find another unplaced short contig (14.5 kb in size) with moderately high Illumina sequencing coverage that contains the 107-bp satellite (**Figure 4 - figure supplement 5**), and this satellite is enriched at the start of our assembly on Muller F (**Figure 4 - figure supplement 1**). High read coverage suggests that the *D. subobscura* genome contains several Mb of the 12-bp and 107-bp satellite sequence, but that these simple satellite arrays are largely collapsed in our assemblies. FISH analysis confirms that the 12-bp repeat is enriched at all the centromeres of *D. suboboscura* chromosomes, except Muller F (and our genomic analysis suggests that the 107-mer is a good candidate for the centromere-associated repeat on Muller F). The intensity of fluorescence varied dramatically across chromosomes, and some chromosomes have exceptionally large foci that overwhelmed the visualization of other chromosomes (**Fig. 4C, Figure 4 - figure supplement 4**). This suggests that the size of the satellite array may differ drastically across different chromosomes.

Thus, our bioinformatics analysis identifies putative centromere-associated repeats in each species (**Table 2**), and FISH supports their location near the centromere. Importantly, however, our experiments do not allow us to identify the functional centromere. Indeed, initial FISH experiments in *D. melanogaster* suggested that the functional centromere in that species consist of simple satellites (Lohe *et al.* 1993; Sun *et al.* 1997; Török *et al.* 2000; Jagannathan *et al.* 2017; Talbert *et al.* 2018). However, recent high-resolution chromatin fiber imaging and mapping of the centromere protein instead indicate that island of retroelements, flanked by large arrays of satellites are the major functional components of centromeres in *D. melanogaster* (Talbert *et al.* 2018; Chang *et al.* 2019). It will be of great interest to perform similar high-resolution mapping studies in the *obscura* group, to determine the functional centromere and its evolution in this species group.

Gene synteny reveals 'genomic accordions' associated with centromere gain and loss

Given the highly repetitive nature of centromeres, their embedment in repetitive pericentromeric DNA, and the fast turn-over of repeats and centromeric satellites between species, studying centromere evolution based on centromere (or pericentric) DNA itself is challenging. Instead, we used protein-coding genes inside or flanking pericentromeric DNA to obtain clues about the evolutionary history of centromere relocation and broad-scale patterns of karyotype evolution.

Figure 3B gives an overview of global syntenic relationships across the species investigated, based on the location of protein-coding genes; genes are assigned to Muller elements (and color-coded accordingly) based on their location in *D. suboobscura*. Consistent with previous studies within the Drosophila genus (Bhutkar *et al.* 2008), syntenic comparisons on all Muller elements reveal a rich history of intrachromosomal reshuffling of genes (**Fig 3B**). Muller C is telocentric in every species, but no major changes in overall chromosome size or repeat structure are found (**Fig 3B**, **Fig. 2**). Likewise, Muller D fused to Muller A in more derived species of the *affinis* or *pseudoobscura* subgroups, but otherwise shows no major changes in size or composition. We detect a rearrangement in *D. athabasca* that moved some Muller A genes inside Muller D (**Fig. 3B**), but the other three species with the Muller A and D fusion (*D. miranda*, *D. pseudoobscura* and *D. lowei*) do not show any reshuffling of genes between Muller A and D. This suggests that this re-arrangement occurred in the *D. athabasca* lineage, and argues against the proposed model of chromosome evolution in this species group (Segarra and Aguadé 1992; Segarra *et al.* 1995). Rather than a fusion between a telocentric Muller A and D followed by a pericentric inversion translocating genetic material from Muller A to Muller D, synteny relationships reveal no evidence of a pericentric inversion in an

ancestor of the *pseudoobscura* and *affinis* subgroups. In addition, analysis of a draft genome of *D. bifasciata* (a member of the *obscura* subgroup that lacks the Muller A-D fusion) further confirms that the centromere position on Muller A is conserved between *obscura*, *pseudoobscura* and *affinis* flies (**Figure 3 - figure supplement 2**). Conservation of centromere position on Muller A corroborates that this chromosome was metacentric in an ancestor before the fusion between Muller A and D occurred.

In contrast, broad genomic comparisons reveal an accordion-like pattern of DNA gain and loss in chromosomes that have undergone a shift in centromere location (**Fig. 3B**). The novel emergence of a centromere inside a chromosome arm results in a dramatic size increase in the genomic region now containing the centromere and pericentromere (that is, Muller A in *D. athabasca*, *D. lowei*, *D. miranda* and *D. pseudoobscura*; Muller B in *D. athabasca*; and Muller E in *D. athabasca* and *D. lowei*). Presumably, the invasion of repetitive elements in these newly formed 'neo-centromeres' has diluted their gene content and resulted in a dramatic increase of their size (see below). Intriguingly, once a centromere has been lost in the center of a chromosome, the genomic region associated with the now defunct 'paleo-centromere' dramatically decreases in size, and gene density recovers to a more typical level (that is Muller B in *D. lowei*, *D. miranda* and *D. pseudoobscura;* and Muller E in *D. miranda* and *D. pseudoobscura*).

Emergence of new centromeres in gene-rich regions

To carefully reconstruct the molecular changes associated with the shift from an telocentric to a metacentric chromosome, we identified genes inside the pericentromere on Muller A, B, and E in species where these chromosomes are metacentric, and identify their homologs in *D. subobscura*, the species with the ancestral karyotype. We focused on Muller A from *D. pseudoobscura*, and Muller B and E from *D. athabasca* (which have 42, 93, and 73 genes inside their pericentromere with a homolog in *D. subobscura*, respectively), but the same conclusions are reached if we select metacentric chromosomes from other species (**Figure 5 - figure supplement 1**). Note that gene annotations can be challenging in highly repetitive pericentromeric regions, likely leading to some mis-annotation. Therefore, we focused our analyses on orthologs identified through reciprocal best BLAST hits between genomes.

Intriguingly, the pericentromeric regions on each of the three metacentric chromosomes are largely syntenic with a small number of euchromatic regions in *D. subobscura* that have a similar number and orientation of genes. This indicates that syntenic relationships are largely conserved and was utilized to define ancestral 'seed regions' of new pericentromeres (that is genomic regions in *D. subobscura* that have become part of the pericentromere on the metacentric Muller A, B or E). In particular, we used reciprocal

best BLAST hits between *D. pseudoobscura* or *D. athabasca* pericentromere genes and *D. subobscura* genes and found the edges of clustered hits (> 2 sequential genes) in the *D. subobscura* genome to define the boundaries of pericentromere seed regions. Note that intrachromosomal rearrangements can break up ancestrally contiguous seed regions or secondarily incorporate additional fragments into a new pericentromere.

As suggested by our broad-scale synteny comparisons (**Fig. 3B**), this more fine-scale look confirms that novel pericentromeres were seeded in previously gene-dense regions on Muller A, B and E, and dramatically extended in size by the accumulation of repetitive DNA (see **Fig. 2, 3**). The pericentromere on *D. pseudoobscura* Muller A is roughly 14.1 Mb (**Table 1**) and contains 42 genes with orthologs in *D. subobscura* (**Fig. 5**). These genes are found in two 'seed regions' on Muller A in *D. subobscura*, whose combined size is 0.3 Mb. Comparison with *D. athabasca* identifies a third syntenic region in the *D. subobscura* genome that became incorporated into the pericentromere of the metacentric Muller A in *D. athabasca* (**Figure 5 - figure supplement 1**). Similarly, the pericentromere of Muller B in *D. athabasca* is roughly 22.8 Mb, and we identify 93 pericentromere genes that are conserved in *D. subobscura* and found in two seed regions with a total size of 0.66 Mb. Finally, the pericentromere on *D. athabasca* Muller E is roughly 11.2 Mb with 73 conserved genes in two seed regions of *D. subobscura* with 0.88 Mb in size. Thus, the combined assembled pericentromers in the metacentric Muller A, B and E are ~48.1 Mb (**Table 1**), while the orthologous regions in *D. subobscura* are only ~1.81 Mb in size (i.e. a total increase over 27-fold). Each of the 'seed' regions has thus expanded dramatically and this expansion appears driven by the accumulation of repetitive sequences (**Fig. 2** and below).

Loss of metacentric location of centromeres

Intriguingly, in some species of the *pseudoobscura* subgroup, the metacentric autosomes secondarily reverted back to telocentric chromosomes (Muller B in all *pseudoobscura* species and Muller E in *D. miranda* and *D. pseudoobscura*). Comparisons of gene content between species reveals that the majority of the genes previously located in the metacentric pericentromere remain on the chromosome arm after this shift in centromere position happened (**Fig. 3B**, **Fig. 5**). That is, the centromere was moved while leaving behind the majority of genes in the pericentromeric region. Syntenic relationships remain mostly intact during this transition, and formerly pericentric genes in *D. athabasca* are found in two genomic segments each on both Muller B and E of *D. pseudoobscura* (**Fig. 5**). However, the size of these former pericentromeric regions (i.e. paleocentromeres) has decreased substantially, through massive purging of intergenic DNA (mainly repetitive elements) while maintaining their gene complement. Specifically, while

the pericentromere of Muller B in *D. athabasca* is 22.8 Mb, most homologous genes in *D. pseudoobscura* are located in a 0.8 Mb sized region in the center of Muller B (58 genes) and almost all remaining genes are found in the pericentromere (22 genes). Likewise, the homologous genes in the 11.2 Mb large pericentromere on *D. athabasca* Muller E are found in two segments along the chromosome in *D. pseudoobscura* (they are roughly 0.96 Mb and 0.18 Mb in size and contain 58 and 12 genes). Similar islands are found in *D. miranda* at homologous positions on both Muller B and E (**Fig 2**; Mahajan *et al.* 2018). In *D. lowei*, the telocentric Muller B shows a similar pattern with the former metacentric pericentromere having drastically contracted in size (to 2.81 Mb). In contrast, Muller E is still metacentric in *D. lowei* (**Fig 2**). This indicates that the transition of a metacentric Muller B to a telocentric chromosome happened in an ancestor of the *pseudoobscura* subgroup, while Muller E only became telocentric after the split from *D. lowei*.

As a mark of their evolutionary past, we find that the repeat content in paleocentromeres is still increased above background levels, despite their massive reduction in size. In *D. pseudoobscura*, for example, the repeat content in euchromatic DNA (that is, non-pericentromere) is ~6%; in the paleocentromeres, ~18% of DNA is masked for repeats on Muller B and ~32% on Muller E (**Fig. 2; Supplementary File 5**). Similar patterns of repeat enrichment are seen for *D. miranda* (**Fig. 2; Supplementary File 5**). In addition, paleocentromere islands on both Muller elements show a clear spike in heterochromatin enrichment (as measured by H3K9me3 profiles, see **Fig. 2**; Mahajan *et al.* 2018). Thus, despite having been purged of a large fraction of repeats, these 'paleocentromeres' still contain clear signatures of their former life as a pericentromeric region, with the younger paleocentromere (on Muller E) showing higher repeat content.

Phylogenetic dating allows us to roughly estimate when these evolutionary transitions occurred and suggests that the birth and death of a centromere can happen quickly on an evolutionary time scale. Molecular dating in this species group suggests that the *subobscura* subgroup split from the other species groups about 20MY ago (Gao *et al.* 2007). The *affinis* and *pseudoobscura* subgroups split from the *obscura* flies about 16MY ago and diverged from each other roughly 14MY ago (Gao *et al.* 2007). *D. lowei* diverged about 9MY ago, and *D. miranda* and *D. pseudoobscura* split roughly 3MY ago (Gao *et al.* 2007). These rough time estimates suggest that Muller A, B and E all became metacentric 16-20MY ago. Muller A and D subsequently fused (14-16MY ago), and Muller B reverted to a telocentric chromosome 9-14MY ago, while Muller E shifted its centromere only 3-9MY ago.

Centromere repositioning drives karyotype differences

What mutational events are creating the diversity of karyotypes across flies from the *obscura* subgroup? Centromeres can shift along the chromosome by either moving the existing centromere to a new location through structural rearrangements, or by *de novo* formation of a centromere at a new genomic position (**Fig. 6A**). Pericentric inversions are typically invoked to explain transitions between metacentric and telocentric chromosomes in flies (Muller 1940; Patterson and Stone 1952). In particular, pericentric inversions where one breakpoint is located within the pericentromere and the other in the euchromatic arm can easily move the centromere along the chromosome, and transform a metacentric chromosome to an telocentric one, and vice versa (**Fig. 6A**). Alternatively, centromeres may emerge *de novo* in a previously euchromatic region by centromere repositioning (**Fig. 6A**). This model involves the *in situ* creation of a neo-centromere without translocation of DNA, concordant with the inactivation of the ancestral centromere. The initial shift in the position of the centromere may be caused by epigenetic changes, resulting in the incorporation of the centromeric histone protein cenH3 at a novel position along the chromosome, and simultaneous silencing of the ancestral centromere (Amor *et al.* 2004; Tolomeo *et al.* 2017; Schubert 2018).

The reconstruction of syntenic relationships between species with different karyotypes should allow us to infer the mutational mechanisms driving the movement of centromeres across the genome, yet the high frequency of inversions in Drosophila makes the exact reconstruction of karyotype evolution difficult. Also, the repetitive nature of centromeres, and their fast turnover between species makes centromere DNA itself unsuitable for distinguishing between models of centromere movement, but unique regions within or adjacent to pericentromeric heterochromatin can be used (**Fig. 6B**). If pericentromeric inversions move the centromere along the chromosome, we expect that some genes are part of the pericentromere in both karyotypic configurations (**Fig. 6B**). In contrast, if centromeres are seeded *de novo* at a new genomic location, we expect no overlap in pericentromeric gene content (or genes immediately flanking the pericentromere) between species with different karyotypes (**Fig. 6B**). Note that we cannot refute more complex structural rearrangements, such as subsequent peri- and paracentric inversions with breaks flanking the core centromere repeat on either side, where the pericentric inversion moves the centromere to a new chromosomal position, followed by a paracentric inversion moving formerly pericentromeric genes to another location (Schubert 2018).

Genome-wide alignments between *D. subobscura* (or *D. guanche*) and *D. athabasca* or species from the *D. pseudoobscura* subgroup are consistent with centromere repositioning creating novel centromeres in the middle of Muller A, B and E from an ancestrally telocentric chromosome (**Fig. 5**). In particular, all the homologs of genes present in the pericentromere region of all three metacentric Muller elements are

463 located in euchromatic, gene-rich regions in *D. subobscura* (**Fig. 5**). This is consistent with a model where
464 the emergence of metacentric Muller A, B and E elements was due to *de novo* origination of a centromere
465 in the middle of these chromosomes (but could also be caused by complex structural re-arrangements).
466
467 Close inspection of the genes adjacent to the centromeres on Muller A, B and E in *D. subobscura* further
468 shows that their homologs are located nowhere near the (peri)centromere in species with metacentric
469 chromosomes (**Fig. 6C**). Instead, *D. subobscura* (peri)centromeric genes on Muller A are found in the middle
470 of the long arm of Muller A in *D. pseudoobscura,* pericentromeric Muller B genes are located in the center
471 of the short arm of Muller B in *D. athabasca,* and pericentromeric Muller E genes in *D. subobscura* are
472 found in the center of the short arm of Muller E in *D. athabasca* (**Fig. 6C**). As mentioned above, an
473 additional genomic fragment that is adjacent to the centromere of *D. subobscura* became incorporated into
474 the pericentromere of Muller A in *D. athabasca* (**Figure 5 - figure supplement 1**), and pericentric genes in *D.*
475 *subobscura* appear in the pericentromere of *D. athabasca* but not near the inferred centromere position
476 (**Figure 5 - figure supplement 1**). Thus, genomic comparisons do not support a simple model where
477 pericentric inversions move the centromere from a telocentric location to the center of the chromosome,
478 as previously proposed. However, we cannot exclude complex rearrangements (such as multiple nested
479 inversions, chromosome breakage, etc.) contributing to centromere shifts in *obscura* flies.
480
481 In species of the *pseudoobscura* subgroup, Muller B and E have become telocentric secondarily. The shift of
482 the centromere on Muller E from a central to a terminal location again shows no evidence for involvement
483 of inversions. Homologs of genes found in the pericentromere of *D. athabasca* (or *D. lowei*) are all located
484 along the euchromatic chromatin arm in both *D. miranda* and *D. pseudoobscura*, mega-bases away from
485 the centromere on Muller E (**Fig. 5**). Concordantly, the homologs of genes adjacent to the centromere on
486 Muller E of *D. miranda* are found on the short arm of Muller E in *D. lowei*, nowhere close to the centromere
487 in this species (**Fig. 6C**).
488
489 In contrast, we find evidence for a pericentric inversion transforming a metacentric Muller B into a
490 telocentric chromosome in the *pseudoobscura* subgroup. In particular, we find that the homologs of genes
491 located within the pericentromere of *D. athabasca* are split in *pseudoobscura* group flies, with most of
492 them remaining in the center of the chromosome (and forming the heterochromatic paleocentromere
493 discussed above). Intriguingly, a second set of genes that is pericentric in *D. athabasca* is also located within
494 the pericentromere of Muller B in *D. pseudoobscura* flies (**Fig. 5**). In addition, genes that flank the
495 telocentric centromere in *D. lowei* also border the metacentric centromere of *D. athabasca* (**Fig. 6C**). Thus,

these syntenic relationships suggest that a pericentric inversion was involved in the transition of a metacentric Muller B to a telocentric chromosome in the *pseudoobcura* group. However, the inferred inversion has left behind the majority of genes from the ancestral pericentromere, which constitute the heterochromatic paleocentromere in *pseudoobscura* flies described above. Thus, while our data do not allow us to rule out complex structural re-arrangements, we infer that simple centromere repositioning events underlie most centromere shifts observed in the *obscura* group, and the emergence of new karyotypes. However, pericentric inversions also contribute to karyotype evolution.

Seed regions for neo-centromeres are AT-rich

What molecular characteristics drive the emergence of a new centromere in a previously euchromatic region? The genomic regions in *D. subobscura* that became incorporated into the pericentromeres in *D. athabasca* and the *pseudoobscura* subgroup (i.e. the pericentromere 'seed' regions) overall look like random, anonymous regions in *D. subobscura* (**Fig. 2, Figure 5 - figure supplement 2**). Both gene density and repeat density in these 'seed' regions appear within the bounds of background levels of gene and repeat density in euchromatic regions in *D. subobscura*. We do detect an overall significant increase in repeat content and gene content (**Supplementary File 7**); however, this appears to be driven largely by properties of individual seeds (e.g, the increased repetitiveness of the Muller E seed close to the pericentromere or the increased gene density in one of the Muller B seeds, **Figure 5 - figure supplement 2**). The most unusual aspect of each of these seeds regions is that they are very AT-rich (**Fig. 5B; Supplementary File 7**). To better understand what may drive this increase in AT content, we separately compared AT content in intergenic, genic and exon regions (**Figure 5 - figure supplement 3**). Interestingly, we find that AT content is significantly increased for each functional category in seed regions, relative to genomic background (**Supplementary File 7**). Thus, certain chromosomal regions may be predisposed to centromere formation, possibly because of their high AT content.

Gene evolution in neo-centromere and paleo-centromere

In primates, newly evolved centromeres are notably deficient in genes, suggesting that centromere function is incompatible with gene activity (Lomiento *et al.* 2008). In contrast, the neo-centromeres in *D. athabasca* and flies of the *pseudoobscura* group appear to have been seeded in previously gene-rich regions on Muller A, B and D (**Fig. 5**). For example, the genomic regions that became part of the centromere harbor a total of 245 genes annotated on Muller A, B and E in *D. athabasca*. The formation of the (peri)centromere is accompanied by a massive gain in repetitive DNA, and a change in the chromatin environment for genes in these regions, with a global increase in H3K9me3 enrichment (**Fig. 2**). Thus, genes

formerly located in euchromatic chromosome arms are now found in pericentromeric heterochromatin. In contrast to the major increase in repetitive DNA that we observed when a centromere is newly gained, we find that the former centromeres (or paleo-centromeres) in *D. miranda* and *D. pseudoobscura* are losing repeats, yet still maintain a heterochromatic environment (**Fig. 2**).

To evaluate the functional significance of the changes in chromatin environment, we contrasted chromatin structure and expression levels of genes across species and genomic regions. We generated metagene plots of H3K9me3 enrichment for different categories of genes and their flanking sequences in *D. subobscura* and *D. athabasca*. As expected, genes located on the chromosome arms outside the pericentromeric region generally show no enrichment of H3K9me3 (**Fig. 7A**). In contrast, the small dot chromosome of Drosophila is heterochromatic yet gene-rich, and genes on this chromosome have adopted a unique chromatin structure (Riddle *et al.* 2012). While there is a depletion of repressive chromatin marks at their transcription start site, H3K9me3 is strongly enriched over gene bodies (Riddle *et al.* 2012), and this unique chromatin signature is regulated by a molecular pathway unique to the dot chromosome. Genes in pericentric heterochromatin of *D. melanogaster*, on the other hand, have a distinct chromatin structure and do not show this enrichment of H3K9me3 at their gene bodies (Riddle *et al.* 2012). Metagene plots for genes located on the dot chromosome for *D. athabasca* and *D. suboobscura* are similar to those in *D. melanogaster* and show a depletion of H3K9me3 at their promoter region, but an enrichment over gene bodies (**Fig. 7A**). Genes in the pericentromere of *D. subobscura* resemble pericentromeric genes in *D. melanogaster*, with rather uniform levels of H3K9me3 across the genes and flanking regions, at slightly higher levels than euchromatic genes (Riddle *et al.* 2011; 2012). Genes in the neo-centromeres of *D. athabasca* show a unique chromatin signature: H3K9me3 levels are depleted at their transcription start site and across their gene body, yet H3K9me3 levels are strikingly elevated at flanking regions (**Fig. 7A**). Thus, while pericentric genes in *D. athabasca* are embedded in repressive heterochromatin, silencing marks do not seem to spread across gene bodies. In line with a lack of repressive chromatin marks, we find no reduction in expression levels of genes within pericentromeric regions of metacentric chromosomes in *D. athabasca* relative to genes located in neighboring euchromatin (in fact, genes are expressed at significantly higher levels inside the pericentromere; **Fig. 7B**). Levels of gene expression between homologous genes in *D. athabasca* and *D. suboobscura* ('seed' vs. pericentromere, **Fig. 7B**), further demonstrate that pericentric genes in *D. athabasca* are not expressed at a lower level, despite being embedded in a repressive heterochromatic environment.

Repeat evolution in pericentromeric regions

We show that newly formed (peri)centromeres dramatically increase in size, due to a massive gain in repetitive elements. To investigate what types of repeats are driving this size increase, and whether similar elements cause size expansion in different chromosomes and species, we categorized repeats from our assembled genomes into different repeat families and elements, using a curated TE library (Hill and Betancourt 2018). We also performed an independent characterization of repetitive elements using *de novo* assembly of repeats with dnaPipeTE (Goubert *et al.* 2015). Overall, we find dramatic differences in repeat density and composition between species (**Table 3**, **Fig. 8, Supplementary File 8**). As mentioned, *D. subobscura* chromosomes have by far the lowest fraction of bases masked for TEs (6%) and only show slight enrichment near the assembled pericentromeres (**Table 3**, **Fig. 8**, **Figure 8 - figure supplement 1**). Genomic repeat content is substantially higher for the other species, ranging from 19-25% (**Table 3**) and these patterns were qualitatively similar in our dnaPipeTE analyses (**Figure 8 - figure supplement 2**). Repeat composition also differs among taxonomic groups. A retrotransposable element belonging to the *Gypsy* family (Dpse_Gypsy_6) is the most prevalent TE in *D. subobscura*, and accounts for 17% of TE-derived DNA in that species (**Table 3**, **Supplementary File 9**). The most common repeat in *D. athabasca* belongs to the Jockey family, with 9% of all TEs derived from the Daff_Jockey_18 element (**Table 3**). In species of the *pseudoobscura* subgroup, on the other hand, a LINE repeat belonging to the CR1 family (CR1-1_DPer) and a helitron element (HelitronN-1_DPe) are the most common repeats, with both repeats accounting for a similar fraction of repeat-masked bases in each genome (6-9% each; see **Table 3**).

As expected, the vast majority of repeats accumulate within pericentromeric regions (**Fig. 8**), but some elements are predominantly found along chromosome arms (**Figure 8 - figure supplement 1**). For example, *Gypsy8* appears enriched in euchromatic arms and several unspecified TEs are only found scattered across euchromatic regions of each chromosome (**Figure 8 - figure supplement 1**). We typically find that the same TE is the dominant transposable element for all pericentromeres within a species, even if they were formed at different time points. Thus, there is more similarity among pericentromeric TE families between chromosomes within a species than there is between homologous chromosomes between species (**Figure 8 - figure supplement 1**). Overall, genomic patterns of TE abundance reflect the dynamic evolution of repeats across the genome and the quick turnover that can occur between species.

Discussion

The centromere is a complex cellular structure responsible for proper chromosome segregation during cell division. In almost all eukaryotes, centromeres are found at specific locations along chromosomes and are composed of (or embedded in) blocks of satellite DNA that occasionally are very large (mega-bases in size). Yet, in spite of the fundamental importance of centromere function and the high conservation of most centromeric proteins (van Hooff *et al.* 2017), both the centromere-specific histone cenH3, and the centromeric satellite DNA sequences can differ substantially even among closely related species (Henikoff *et al.* 2001; Melters *et al.* 2013; Steiner and Henikoff 2015). Further, centromere size and position can change rapidly along the chromosome.

Classical and molecular cytogenetic studies have provided rich data on the karyotypes of thousands of species in the last century (including information on diploid numbers, and relative length and morphology of chromosomes), and have revealed a dynamic evolutionary history of chromosomes across eukaryotes. The different positions of centromeres along a chromosome were historically almost always interpreted as the result of pericentric inversions or complex rearrangements. However, comparative cytogenetic studies of chromosome synteny among species, within a phylogenetic context, have revealed that centromere location on homologous chromosomes may change with no concomitant change in DNA marker order. 'Centromere repositioning', that is, the movement of the centromere along the chromosome without marker order variation, was first given as a possible mechanism for karyotype evolution in a broad investigation of chromosomal evolution in 60 species of primates (Dutrillaux 1979). Over the past few decades, numerous studies have found evolutionary new centromeres in different lineages, including fungi, insects, birds, and mammals (O'Neill *et al.* 2004; Marshall *et al.* 2008; Rocchi *et al.* 2012; Scott and Sullivan 2014; Schneider *et al.* 2016; Burrack *et al.* 2016; Tolomeo *et al.* 2017). Some of the observed karyotype reshuffling may be due to the inheritance of neo-centromeres (Amor *et al.* 2004) while some may be the product of successive pericentric inversions (Rocchi *et al.* 2012). However, almost all studies have used cytological techniques with limited resolution to track karyotype evolution at the DNA sequence level.

Drosophila, the workhorse of classical and molecular genetics, has a long history and rich literature of karyotype evolution (Muller 1940; Patterson and Stone 1952). Inversions are ubiquitous in flies. Thus, while dozens of transitions between telocentric and metacentric chromosomes have been documented in Drosophila (Muller 1940; Patterson and Stone 1952), pericentric inversions are typically invoked to explain the movement of centromeres between species (Segarra *et al.* 1995; Schaeffer *et al.* 2008). Here, we reconstruct karyotype evolution in flies using modern genomic techniques. Contrary to the common belief that pericentric inversions are the main mechanism for reshuffling karyotypes among different Drosophila

lineages, our analysis suggests that centromere repositioning appears to play an important role in chromosome evolution in Drosophila. Note however, that we cannot refute complex structural rearrangements having contributed to the movement of centromeres across species.

Neo-centromeres and *de novo* formation of centromeres

The location of the centromere on a given chromosome is inherited, such that syntenic centromeric loci are found in related species. Yet, centromeres are thought to be specified epigenetically, and can form at many positions along a chromosome (Karpen and Allshire 1997; Allshire and Karpen 2008). Neo-centromeres occur when kinetochores assemble *de novo*, at DNA loci not previously associated with kinetochore proteins. They are sometimes found at acentric chromosome fragments that arise by chromosome rearrangements, deletions, or amplifications and can restore the ability of an acentric chromosome fragment to segregate efficiently (Burrack *et al.* 2016). In rare cases, neo-centromeres form in otherwise normal chromosomes, without physical deletion of the native centromere, presumably following inactivation of the native centromere through unknown mechanisms (Amor *et al.* 2004; Liehr *et al.* 2010). More than 100 human neo-centromere locations have been identified (Marshall *et al.* 2008), and neo-centromere formation has been studied in several model systems including Drosophila (Maggert and Karpen 2001), fungi (Ishii *et al.* 2008; Burrack *et al.* 2016) or chicken cell lines (Shang *et al.* 2013). Detailed analysis of neo-centromeres in humans and other species has shown that their formation is a rare and usually deleterious event, yet neo-centromeres can occur at many different regions across the genome. Sites of neo-centromere formation do not share a unique DNA sequence but may have in common other features that predispose to neo-centromere formation, such as increased AT content (Amor and Choo 2002; Amor *et al.* 2004; Marshall *et al.* 2008). Fixations of neo-centromeres within a species result in the formation of 'evolutionary new centromeres'. Evolutionary new centromeres can be relatively frequent in some groups. In macaque, for instance, nine out of 20 autosomal centromeres are evolutionarily new; in donkey at least five such neo-centromeres originated after divergence from the zebra, in less than 1 million years. The establishment of new centromeres is often associated with an accompanying expansion of satellites at the new centromere location and loss of large satellite arrays at the former location (Ventura *et al.* 2007). Although most human neo-centromeres are pathological chromosome aberrations that are evolutionarily rare and transient, it has become apparent that on occasion, the shift of a centromere to a new position becomes fixed, resulting in a permanent karyotypic change (Cardone *et al.* 2007).

Thus, detailed functional studies on neo-centromeres combined with comparative cytogenetic investigations of evolutionary new centromeres support that centromere repositioning may be important

for karyotype evolution. As noted, however, investigations of these phenomena at the DNA level across species are lacking. Despite dramatic advances in sequencing technology, the repeat structure of centromeres has banished them to the last frontier of the genome to be deciphered for most organisms. Our study is the first to track the evolution of the centromere in flies at the sequence level, using high-quality genomic assemblies. Our comparative analysis allows us to reconstruct the entire life cycle of complex centromeres, including their formation, maturation and degradation.

Seeding of new centromeres

We find that new centromeres in flies of the *obscura* group were formed in initially anonymous regions (**Fig. 5**). The most striking feature of seed regions is their increased AT-content (**Fig. 5, Figure 5 - figure supplement 2**), which resembles observations from evolutionary novel centromeres in primates (Federico *et al.* 2017), or neo-centromeres in humans (Amor and Choo 2002; Amor *et al.* 2004; Marshall *et al.* 2008) or chicken (Shang *et al.* 2013). Such (neo)centromeres typically arise in euchromatic regions that are otherwise unremarkable with respect to sequence composition but are reported to predominantly emerge in ancestral regions that are gene-poor and AT-rich. The lack of any obvious sequence features involved in the formation of new centromeres confirms the notion that the origin of a neo-centromere appears to be associated with epigenetic phenomena that are not yet well understood, and increased AT content may provide a more favorable disposition for the formation of novel centromeres. In other respects, these seed regions show no particular differences from bulk DNA sequences. In contrast to (neo)centromere formation in humans or other model systems, new centromeres in flies are seeded in ancestral regions that show typical gene density. Thus, centromere function in flies is not incompatible with gene activity, and we find that genes embedded in pericentromeric heterochromatin show normal levels of transcriptional output. This is surprising, since flanking regions of pericentromeric genes show strong enrichment for silencing chromatin marks, yet these epigenetic silencing marks do not spread over promoter regions or gene bodies in neo-centromeres in *D. athabasca*.

Maturation of centromeres

Mature eukaryotic centromeres are composed of arrays of satellite DNA frequently surrounded by clusters of transposable element islands (She *et al.* 2004), or the other way round (Talbert *et al.* 2018; Chang *et al.* 2019), and are embedded in silencing heterochromatin. The vast majority of evolutionary new centromeres that have been investigated in primates possess a heterochromatic block similar to normal centromeres. FISH studies in macaque, for instance, have shown that all the nine novel centromeres formed in this lineage have large blocks of alpha-satellite DNA indistinguishable from other macaque centromeres

693 (Ventura *et al.* 2007). Detailed analysis of one of these new centromeres has shown that it was seeded in a
694 gene dessert, suggesting that the absence of genes may have been a prerequisite for survival of this
695 centromere (Ventura *et al.* 2007). However, the molecular nature and sequence composition of new
696 centromeres is generally unknown.
697
698 Our study is the first that bridges this gap, and we can directly infer the molecular and epigenetic changes
699 accompanying the maturation of centromere. In contrast to findings in macaques, the newly formed
700 centromeres in *Drosophila* species emerged in gene-rich sequences. We further demonstrate that their
701 establishment was accompanied by the acquisition of species-specific arrays of centromere-associated
702 satellite DNA as well as clusters of pericentromeric transposable elements. In particular, the seeding
703 domains for new centromeres in *D. subobscura* show a slight increase in both gene and repeat density but
704 are devoid of arrays of satellite DNA. Following the seeding of a novel centromere on Muller A, B and E,
705 these regions expanded dramatically in size. Most of this size increase appears to be driven by an
706 accumulation of transposable elements. Different TEs accumulate at pericentromeres in different fly
707 species, but we typically find that the same element as the dominant TE for all centromeres within a
708 species. Thus, genomic patterns of TE abundance reflect the dynamic evolution of repeats.
709
710 Interestingly, we find taxon-specific satellites in each group that may represent the centromeric repeat. In
711 our assemblies, we recovered a shared putative satellite repeat in *D. miranda*, *D. pseudoobscura* and *D.*
712 *lowei* (a 21mer) and a 99mer unique to *D. miranda*, and a different 160bp satellite in *D. athabasca*. Each of
713 these satellites was enriched in the (peri)centric region of most chromosomes in our assemblies, suggesting
714 that our genome assemblies may cover at least part of the centromeric satellite, and localization of these
715 satellites near the centromere was confirmed by FISH. Only in *D. subobscura*, the species with the ancestral
716 karyotype and putatively older (peri)centromeres, we failed to assemble a highly abundant repeat motif as
717 a candidate for the centromeric satellite. Inspection of raw reads identified a highly abundant 12mer that
718 was collapsed in our assembly but localized to the centromere in this species. Thus, long-read sequencing
719 data allowed us to assembly (parts of the) younger, more heterogeneous pericentromeres, but not the
720 older pericentromere of *D. subobscura*, which may be more homogenous (i.e. the satellite repeats may be
721 interrupted by fewer TEs; see below). Note, however, that additional experiments are necessary to confirm
722 the functional centromere in each species (Talbert et al. 2018; Chang et al. 2019).
723
724 Our data are thus consistent with a model where novel centromeres were formed in regions devoid of
725 repeats, which began to accumulate TEs after their formation. TEs might provide the substrate for satellite

DNA formation and trigger the initiation of species-specific satellites. Younger centromeres may be more heterogeneous and be a conglomerate of different TEs and satellites, and occasionally be interrupted by single-copy genes (and may resemble the younger pericentromeres observed in the *affinis* and *pseudoobscura* group). Homogenizing forces, such as gene conversion, could then operate to erode heterogeneity and complexity of satellites, leading to the evolution of simple and highly homogeneous satellite arrays observed in many old centromeres (as possibly found in *D. subobscura*).

Epigenetic models of centromere specification stress the importance of cenH3 nucleation in formation of a centromere (Karpen and Allshire 1997; Allshire and Karpen 2008). We show that the region that became part of the mature pericentromere is not limited to a single nucleation site (that is, encompassing only a few nucleosomes). Instead, the seed region spreads over several 100-kb for each of the newly formed pericentromeres, containing dozens of genes each. In the mature pericentromere, these regions are now many Mb in size due to invasion of TEs, and the genes are now embedded in large stretches of repeats. Silencing heterochromatin is found around many ancestral repeat-rich centromeres, and genes mapping to seed regions are at risk of silencing by successive heterochromatization. Indeed, we find that silencing heterochromatin has spread across the repetitive new pericentromeres and formerly euchromatic genes are now embedded in pericentromeric heterochromatin. However, silencing marks are absent at promoter regions, and do not seem to spread across gene bodies despite drastic enrichment at intergenic regions directly flanking these pericentromeric genes. In line with a lack of repressive chromatin marks, we find no reduction in expression levels of genes within pericentromeric regions. How spreading of heterochromatin across gene bodies is prevented is not clear and will require future detailed experimental investigation.

Centromere deactivation

The kinetochore is formed on the centromere, and the presence of multiple centromeres on a single chromosome can cause abnormal spindle-chromosome interaction resulting in chromosome missegregation or chromosome breakage during cell division (Brock and Bloom 1994). Having one and only one centromere per chromosome is thus essential for proper chromosome segregation. Centromere repositioning creates a novel centromere, suggesting that a single chromosome contains multiple latent centromeres. Chromosomes with multiple centromeres, such as dicentric chromosomes, can be stabilized by centromere inactivation, a process which reestablishes monocentric chromosomes (Sato *et al.* 2012). Studies of artificial dicentric centromeres in fission yeast demonstrated that the presence of a dicentric chromosome causes cell-cycle arrest in interphase. Epigenetic centromere inactivation could stabilize dicentric chromosomes, and consequent heterochromatinization from pericentric repeats to the central

759 domain could prevent reactivation of the inactivated centromere (Sato *et al.* 2012). However, little is
760 known about this process in naturally occurring dicentric chromosomes (Cech and Peichel 2016).
761
762 Our high-quality genome sequences allow us to recover the remnants of former centromeres in flies in the
763 *pseudoobscura* group, and we can reconstruct the stepwise disassembling of a repeat-rich centromere. In
764 flies of the *pseudoobscura* group, both Muller B and E reverted to telocentric chromosomes, and
765 phylogenetic dating suggests that this transition occurred 9-14 MY ago for Muller B, and 3-9 MY for Muller
766 E. A pericentric inversion appears to have moved a small part of the pericentromere on Muller B, and a new
767 centromere was probably created on Muller E, leaving behind most or all of the metacentric centromere
768 (**Fig. 5**). Our sequence data demonstrate how redundant centromeres can become inactivated and lost
769 rapidly. We detect major re-arrangements at these former centromeres, and both have shrunken
770 dramatically in size, through purging of repetitive DNA. The repeat content of both paleocentromeres is
771 elevated only slightly compared to background regions. Interestingly, however, we find that both
772 paleocentromeres are still strongly enriched for heterochromatin (**Fig. 5**). In fission yeast, inactive
773 centromeres are prevented from reactivation by heterochromatinization (Sato *et al.* 2012), and it is
774 possible that heterochromatin formation at paleocentromeres in *pseudoobscura* flies is necessary to inhibit
775 the assembly of a functional centromere. This is accompanied by removal of centromere-specific
776 sequences, and we detect both rapid turnover of centromere-associated satellites between species, and no
777 satellite enrichment at the decaying paleocentromeres on Muller B and E.
778
779 Dicentric chromosomes can result from centromere repositioning events, or chromosome fusions. As
780 demonstrated by the extensive variation in chromosome number among species, chromosome fusion (and
781 fission) events are common during evolution. For instance, in multicellular eukaryotes, chromosome
782 number ranges from a single pair of chromosomes per cell in the Jack jumper ant (Crosland and Crozier
783 1986) to 630 pairs of chromosomes in the Adders tongue fern (Khandelwal 2008). Chromosome numbers
784 can also vary substantially among closely related species (Tree of Sex Consortium 2014). The diversity of
785 karyotypes across eukaryotes suggests that dicentric chromosomes accompanied by inactivation of one
786 centromere may be common and may leave distinct footprints in the genomic make-up of a species. In our
787 initial analysis of the *D. miranda* genome, we noticed the occurrence of repeat islands in this species, but
788 their significance was unclear (Mahajan *et al.* 2018). Our comparative analysis now clearly demonstrates
789 that this unusual sequence feature was in fact the legacy of an inactivated centromere, and repeat-islands
790 in other genomes may likewise present remnants of former centromeres. Our study thus highlights that a
791 full understanding of genome structure and function is only possibly in an evolutionary framework.

What drives centromere relocation?

Reasons for centromere relocation could be manifold. The function of the ancestral centromere could become compromised by mutations, such as a loss of centromeric satellites, or insertions of transposable elements in the centromere (Roizès 2006). This could result in the activation of a latent novel centromere, and inactivation of the ancestral one. Support for this model comes from investigations of neo-centromeres in humans, which were found on a chromosome with a substantially reduced size of the alpha satellite array on its ancestral chromosome compared to its wildtype size (Hasson *et al.* 2011). Novel centromeres could also become fixed by meiotic drive. Female meiosis is inherently asymmetric: only one of four meiotic products is transmitted to the next generation via the oocyte. Selfish genetic elements can exploit this asymmetry to preferentially transmit themselves, and a novel centromere that outcompetes the ancestral one may invade a population (Henikoff *et al.* 2001). In maize, non-centromeric blocks of satellite-DNA, or ''knobs,'' can transmit themselves more frequently than expected through asymmetric female meiosis, but not symmetric male meiosis. Knob-containing Abnormal 10 chromosomes can outcompete Normal 10 chromosomes and sponsor drive of other knob-containing chromosomes that are incapable of driving on their own (Rhoades and Dempsey 1966). Invasion of such a drive system could lead to the simultaneous replacement of centromeres (or centromere satellites) on multiple chromosomes. Finally, *de novo* centromere formation may be linked to chromosome breakage. CENP-A was shown to be recruited to DNA double-strand breaks (Zeitlin *et al.* 2009) and a neo-centromere could emerge because of the presence of CENP-A at the breakpoint. We find that new centromeres are predominantly seeded in AT-rich regions, and several studies have suggested that AT-rich sequences are prone to breakage (Zhang and Freudenreich 2007; Kato *et al.* 2012). This could provide a link between neo-centromere seeding and AT content through elevated rates of chromosome breakage.

Evolution of centromere-associated satellites

Rapid evolution of centromeric satellite sequences has been observed across metazoan lineages. This rapid evolution is attributed to processes such as molecular drive, leading to the homogenization and fixation of a variant (or subset of variants) across a repeat array (Dover 1982; Dover *et al.* 1982), and both genetic conflict and centromere drive (Henikoff *et al.* 2001), leading to rapid diversification of repeat families between species. Our high-quality genome assemblies combined with bioinformatics approaches to detect satellite DNA allowed us to identify centromere-associated repeats in several species. Intriguingly, we find that the major satellite repeat can change rapidly between closely related species. For example, we find a 99-bp satellite in *D. miranda* that is (mostly) absent in *D. pseudoobscura* or *D. lowei*, suggesting that this

sequence invaded the *D. miranda* genome in the last 1-2 MY ago and became the dominant centromere-associated satellite. *D. guanche* has a 290-bp satellite at its centromere, which is rare and not associated with the centromere in *D. subobscura* (Bachmann *et al.* 2009; Puerma *et al.* 2018; Karageorgiou *et al.* 2019). Thus, new satellites can form rapidly *de novo* and quickly become the dominant repeat within a species.

How do species-specific satellites emerge? Models of satellite evolution suggest two stages in the emergence of new satellites: Amplification processes generate small tandem sequences, and some of these sequences expand to large arrays by unequal exchange (Stephan and Cho 1994). Several studies have found that TEs may contribute to the formation of novel satellite arrays. A stunning example of the birth of a satellite from a TE has been shown in *D. virilis* (Dias *et al.* 2014), where a foldback transposon provided internal tandem repeats that acted as "seeds" for the generation and amplification of satellite DNA arrays. Furthermore, TEs from *D. melanogaster* were found to commonly form tandem dimers, which have the potential to provide source material for future expansion into satellite arrays (McGurk and Barbash 2018). Thus, the genesis of satellites from TE insertions offers a possible explanation for the origination of species-specific centromere satellites. Transposons may start to accumulate on a new centromere and provide a substrate for the formation of satellite arrays either through their internal repeats structure, or by double insertion of a transposon into the same site, creating tandem TEs. Such tandem dimers can then expand into large arrays by recombination events. Indeed, many copies of the 99-bp satellite repeat identified in *D. miranda* show similarity to transposable elements in the *Repbase* repeat library, and future detailed study will help to decipher the evolutionary history of species-specific satellites in this species group.

Our data demonstrate that repetitive sequences can turn over rapidly between species and trigger major genomic responses. Recent data from hybrid model systems have shown that incompatibilities between genomes can be complex, and include interactions among heterochromatin, repeats or repetitive centromeres. It will be of great interest to establish the role rapidly evolving centromeres play in the establishment of genomic incompatibilities among closely related taxa and in the formation of new species (Brown and O'Neill 2010).

854 METHODS
855
856 **Key Resource Table**
857

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
strain, strain background (*Drosophila subobscura* male and female)	14011-0131.10	National Drosophila Species Stock Center (Cornell University)	stock center number: 14011-0131.10	
strain, strain background (*Drosophila pseudoobscura* male and female)	MV2-25	National Drosophila Species Stock Center (Cornell University)	stock center number: 14011-0121.94	
biological sample (*Drosophila lowei*)	Jillo6	isofemale line (deceased)		
biological sample (*Drosophila athabasca* EA)	PA60	isofemale line (deceased)		
biological sample (*Drosophila athabasca* EB)	NJ28	isofemale line (deceased)		
commercial assay or kit	TruSeq Stranded RNA kit	Illumina	cat # 20020595	
commercial assay or kit	TruSeq DNA Nano Prep kit	Illumina	cat # 20015965	
commercial assay or kit	DNeasy Kit	Qiagen	cat # 69504	
commercial assay or kit	Blood & Cell Culture DNA Midi Kit	Qiagen	cat # 13343	
commercial assay or kit	Gentra Puregene Tissue Kit	Qiagen	cat # 158667	
commercial assay or kit	Ligation sequencing kit	Nanopore	SQK-LSK108	
commercial assay or kit	Rapid sequencing kit	Nanopore	SQK-RAD004	
commercial assay or kit	Quick DNA plus Midi kit	Zymo	cat # D4075	
commercial assay or kit	Ligation sequencing kit	Nanopore	SQK-LSK109	
commercial assay or kit	SMARTer Universal Low Input DNA-seq kit	Takara/Rubicon Bio	R400676	
chemical compound, drug	H3K9me3 antibody	Diagenode		
software, algorithm	canu	Koren *et al.* 2017		
software, algorithm	MUMmer	Kurtz *et al.* 2004		
software, algorithm	WTDBG2	Ruan and Li 2019		
software, algorithm	minimap2	Li 2019		
software, algorithm	Bandage	Wick *et al.* 2015		
software, algorithm	BWA	Li and Durbin 2009		
software, algorithm	SAMtools	Li et al. 2009		
software, algorithm	bedtools	Quinlan and Hall 2010		
software, algorithm	QUIVER	Chin *et al.* 2013		
software, algorithm	PILON	Walker et al. 2014		
software, algorithm	RACON	Vaser et al. 2017		
software, algorithm	Juicebox	Durand et al. 2016a		
software, algorithm	Juicer	Durand *et al.* 2016b		
software, algorithm	3D-DNA	Dudchenko *et al.* 2017		
software, algorithm	GATK UnifiedGenotyper	DePristo *et al.* 2011		
software, algorithm	REPdenovo	Chu *et al.* 2016		
software, algorithm	RepeatMasker	Smith *et al.* 1996-2005		

software, algorithm	MAKER	Campbell *et al.* 2014		
software, algorithm	HiSat2	Kim *et al.* 2015		
software, algorithm	StringTie	Pertea *et al.* 2015		
software, algorithm	BUSCO	Simão *et al.* 2015		
software, algorithm	Tandem Repeat Finder	Benson 1999		
software, algorithm	k-Seek	Wei et al. 2018; https://github.com/weikevinhc/k-seek		
software, algorithm	TideHunter	Gao et al. 2019		

Long read sequencing and contig assembly

For each Drosophila genome, we followed a similar assembly approach to (Mahajan *et al.* 2018). Briefly, long-read sequencing data were used to generate sequence contigs, Illumina data were used to polish the contigs and identify non-target contigs (contamination), and Hi-C data were used to place contigs into chromosomal scaffolds. However, given differences in the long-read datatype (PacBio or Nanopore), the amount of polymorphism present in the strains sequenced, the sex of the individuals used for sequencing, and differences in the genome architecture (such as repeat content) across species, slightly different species-specific adjustments were done during each assembly, to maximize the quality of each Drosophila genome. Below we give a brief summary of the assembly approach taken for each species. A detailed description of species-specific adjustments for each assembly and validation of our assemblies are given in Supplementary data 1.

Drosophila athabasca: We extracted high molecular weight DNA from ~100 females of semispecies EA (isofemale line PA60) and EB (isofemale line NJ28) (Wong Miller *et al.* 2017) using a QIAGEN Gentra Puregene Tissue Kit. DNA fragments > 100 kb (estimated from pulsed-field gel electrophoresis) were then sequenced using the PacBio RS II platform with 13 SMRT cells for each semispecies. This resulted in a total of 622,136 reads with and N50 read length of 17,517 for EB and 1,195,701 reads with an N50 read length of 12,314 for EA. PacBio reads were assembled using canu (version 1.6) (Koren *et al.* 2017) with a correctedErrorRate of 0.075 for both the EA and EB assembly. During contig assembly with canu (here and below) we explored various correctedErrorRates spanning recommendations to maximize contig length and minimize misassembly.

Drosophila pseudoobscura: We sequenced males from the reference genome strain (Drosophila Species Stock Number: 14011-0121.94; MV2-25) using five Nanopore MinIOn flowcells (versions 9.4 and 9.5) with the ligation sequencing kit (SQK-LSK108) and the rapid sequencing kit (SQK-RAD004). We combined the total output from these five runs, 1,250,165 filtered reads with an N50 read length of 18,209 with published female derived Nanopore reads of the same line (Miller *et al.* 2018). We were particularly interested in assembling Y chromosome contigs and thus took an approach to enrich our assembly of heterochromatic regions of the genome similar to (Chang and Larracuente 2019). We sequentially assembled the autosomes, the X (Muller A-AD), and the Y chromosome, each time subtracting out reads that mapped uniquely to that specific genomic partition. Remaining reads were used in subsequent assemblies. Nanopore reads were mapped to assemblies using minimap2 (Li 2018) and autosomal and X-linked contigs were identified via comparison with *D. miranda* (Mahajan *et al.* 2018) using MUMmer (Kurtz *et al.* 2004). To first assemble the autosomes, all reads > 10kb were selected using FiltLong (https://github.com/rrwick/Filtlong) and assembled using canu (version 1.7), with the -fast option and a correctedErrorRate of 0.144. The X-linked contigs were assembled using identical settings. The Y-linked assembly was performed using only male derived reads and a correctedErrorRate of 0.166. To distinguish between putative Y-linked contigs and unplaced autosomal contigs in the final assembly, contigs with > 1.5 male/female mean Illumina coverage (described below) were considered putatively Y-linked, with the remaining left unplaced.

Drosophila subobscura: We obtained the isofemale line 14011-0131.10 from the National Drosophila Species Stock Center (Cornell University). We extracted high molecular weight DNA form ~100 females using a Zymo quick DNA plus Midi kit (cat # D4075) following the manufacturers recommendations. Isolated DNA was then size selected for fragments >15 kb using BluePippin (Sage Science) and was sequenced on the Nanopore MinIOn using the SQK LSK-109 ligation protocol and two flow cells (version R 9.4.1). This resulted in a total of 1,689,885 filtered reads with an N50 read length for the combined reads of 17,390. Initial assemblies of our data following methods outlined above resulted in many large misassembled contigs likely due to an observed enrichment of chimeric Nanopore reads in our data. To deal with this issue we assembled reads using canu (version 1.8) and modified several parameters to increase stringency (correctedErrorRate=0.065 corMinCoverage=8 batOptions="-dg 3 -db 3 -dr 1 -ca 500 -cp 50" trimReadsCoverage=4 trimReadsOverlap=500 genomeSize=150m). The resulting error-corrected reads were then assembled into contigs using the WTDBG2 assembler (Ruan and Li 2019) with default settings.

Drosophila lowei: Flies were collected from Woodland Park, CO in August of 2018 (elev. 8,500 ft.) to establish multiple isofemale lines. Due to difficulties in maintaining this species that would allow us to inbreed and reduce heterozygosity prior to sequencing, we extracted DNA from a pool of ~ 60 F1 males from the most productive isofemale line, Jillo6. We extracted high molecular weight DNA using a Qiagen Blood & Cell Culture DNA Midi Kit (cat # 13343) and manufactures recommendations. The isolated DNA was size selected as above for fragments > 30 kb and was sequenced as above using the Nanopore MinIOn SQK LSK-109 protocol, on one R 9.4.1 flow cell. One sequencing run produced a total of 437,080 filtered reads with an N50 read length of 36,073. To assemble the reads into contigs, we followed the methods outlined above for *D. subobscura*.

Short read Illumina data

To aid in genome assembly steps outlined below, we generated Illumina short-read male and female whole genome sequence data for each species. We extracted DNA from a single male and female using a Qiagen DNeasy kit following manufacturers recommendations. DNA libraries were then prepared using the Illumina TruSeq Nano Prep kit and sequenced on a Hiseq 4000 with 100bp PE reads. Plots of Illumina short read data were plotted using *circlize* (Gu *et al.* 2014)

Identification of sex chromosome and non-target contigs

To identify sex chromosome contigs during assembly and flag putative non-target contaminant contigs (e.g., yeast, bacteria) for removal from our assemblies we interrogated each assembly graph output (.gfa file) from our canu assemblies using multiple metrics that were visualized in Bandage (Wick *et al.* 2015). These metrics included the mean, median and standard deviation of male and female Illumina coverage by contig, the ratio of female/male coverage, the median long-read coverage (PacBio or Nanopore) and the GC-content (%). We also compared the connectivity of contigs in the assembly graph since non-target contigs are typically isolated from the target species graph. We also identified the top two BLAST hits for all contigs < 1MB. To estimate Illumina coverage-based metrics (above), short and long reads were mapped to each assembly prior to visualization using BWA MEM (Li and Durbin 2009). Downstream files were processed with SAMtools (Li *et al.* 2009) and coverage estimated using bedtools genomecov (Quinlan and Hall 2010).

Genome polishing

The *D. athabasaca* EA and EB PacBio-based assemblies were individually polished using the SMRT analysis software pipeline and two rounds of QUIVER (Chin *et al.* 2013) and two rounds of PILON (version 1.22) (Walker *et al.* 2014). EA and EB were polished with 47× and 48× coverage of 100bp PE Illumina reads, respectively. The *D. pseudoobscura, D. lowei* and *D. subobscura* Nanopore-based assemblies were polished

using three rounds of RACON (version v1.2.1) (Vaser *et al.* 2017) followed by two rounds of PILON using 115×, 47× and 40× mean coverage of Illumina reads, respectively.

Hi-C libraries and genome scaffolding

To scaffold together contigs from our de novo genome assemblies (above) we used chromatin conformation capture to generate Hi-C data. The Hi-C libraries for *D. athabasca* EA and *D. pseudoobscura* were prepared following methods in (Lieberman-Aiden *et al.* 2009) as described in (Mahajan *et al.* 2018) using the restriction enzyme HindIII. Tissues used for Hi-C were from female third instar larvae for *D. athabasca* and male third instar larvae for *D. pseudoobscura*. For *D. subobscura* and *D. lowei*, *in situ* DNase Hi-C was performed on adult females and males, respectively, as described in (Ramani *et al.* 2016) with a few modifications. Briefly, tissue was ground with a micropestle in 1.5 ml microtube with 250 μl PBS buffer and we rinsed the pestle with 750 μl of PBS buffer. The resulting 1 ml homogenate was mixed gently and passed through 100 μm Nylon mesh (sterile cell strainer) to remove debris. The homogenate was centrifuged at 2000Xg for 5 min at 4°C and the cell pellet fixed with 1 ml of 4% formaldehyde at RT for 10 min. Gently inverting tubes every 2 min. The resulting cells were treated as per (Ramani *et al.* 2016) from steps 4 to 74. DNA Libraries were prepared with an Illumina TruSeq Nano library Prep Kit and sequenced on a Hiseq 4000 with 100bp PE reads.

To identify Hi-C contacts for scaffolding our genome assemblies we used Juicer (Durand *et al.* 2016b) to map reads and generate Hi-C contact maps to orient contigs based on 3D interactions (**Supplementary File 10**). We then used the 3D-DNA pipeline (Dudchenko *et al.* 2017) to scaffold contigs and the resulting hic and assembly files were visualized using Juicebox (Durand *et al.* 2016a). To confirm scaffolding quality and to correct potential mistakes we first compared chromosome assignments from 3D-DNA-scaffolded chromosomes to other closely related species with chromosome-level assemblies using MUMmer (i.e., *D. miranda* and *D. guanche*; Mahajan *et al.* 2018; Puerma *et al.* 2018). Second, we compared female and male whole genome sequencing coverage across 3D-DNA-scaffolded chromosomes to confirm predicted patterns of coverage for sex-linked chromosomes and autosomes. Finalized assemblies were scaffolded together with 3D-DNA scripts and contigs joined with 500 N's (**Fig. S3**).

Hybrid *D. athabasca* genome assembly

Residual heterozygosity in our EA and EB isofemale lines initially led to fragmented assemblies of particular chromosomes. Therefore, we created a hybrid of the two *D. athabasca* assemblies that combines the most contiguous Muller elements from each into one final assembly. Initially, Muller C (which is fixed as a neo-sex chromosome in EB) was highly fragmented in our superior EB assembly and Muller E was highly fragmented in our less contiguous EA assembly. Therefore, we combined the Muller C assembly from EA with the rest of the genome for EB. To make Muller C more 'EB-like' we mapped our female Illumina data (above), called variants using GATK's UnifiedGenotyper (DePristo *et al.* 2011), and used FastaAlternateReferenceMaker to replace EA variants with EB variants. All three assemblies are made available.

Identifying repeats and masking genomes

We used REPdenovo (Chu *et al.* 2016) to annotate repeats in each of our draft genome assemblies prior to gene annotation. We first ran REPdenovo on our raw Illumina sequencing reads derived from the same sex as the genome assembly using the following parameters, MIN_REPEAT_FREQ 4, RANGE_ASM_FREQ_DEC 2, RANGE_ASM_FREQ_GAP 0.8, K_MIN 30, K_MAX 50, K_INC 10, K_DFT 30, ASM_NODE_LENGTH_OFFSET -1, MIN_CONTIG_LENGTH 100, IS_DUPLICATE_REPEATS 0.85, COV_DIFF_CUTOFF 0.5, MIN_SUPPORT_PAIRS 20, MIN_FULLY_MAP_RATIO 0.2, TR_SIMILARITY 0.85, RM_CTN_CUTOFF 0.9 ,with READ_DEPTH and GENOME_LENGTH settings for each species specified based on preliminary genome assemblies and mapping based coverage estimates (above). We then used TBLASTN with the parameters -evalue 1e-6, -

numalignments 1, and -numdescriptions 1 to BLAST translated *D. pseudoobscura* genes (r3.04) to our REPdenovo repeat library and eliminated any hits. For each species, we concatenated our REPdenovo repeats with the Repbase *Drosophila* repeat library (downloaded March 22, 2016, from www.girinst.org) and then used this concatenated fasta file to mask each genome with RepeatMasker version 4.0.7 (Smith *et al.* 1996-2005) using the -no_is and -nolow flags. To characterize specific classes of repeats and their distibutions across the genome we used a curated repeat library for obscura group flies (Hill and Betancourt 2018). Plots were generated using KaryoploteR (Gel and Serra 2017).

Genome annotation

To annotate our genome assemblies we used the MAKER annotation pipeline (Campbell *et al.* 2014). For *D. athabasca* and *D. pseudoobscura*, we used gene expression data (RNA-seq) to generate transcriptomes which were then used as ESTs during annotation. For *D. athabasca* we first extracted total RNA using the TRIZOL RNA isolation method and generated RNA-seq data using the Illumina TruSeq Stranded RNA kit. RNA was isolated for multiple tissues and developmental stages from both EA and EB. Individual RNA-seq libraries consisted of pools of tissue from multiple individuals and included ovaries, testes, male and female adult heads, male and female whole larvae, and male and female larval heads (**Supplementary File 1**). All RNA-seq libraries were sequenced using an Illumina HiSeq 4000 with 100bp PE reads. Raw Illumina reads were aligned to the genome using HiSat2 with default settings and the -dta flag for transcriptome assembly (Kim *et al.* 2015). We then used StringTie (Pertea *et al.* 2015) to build semispecies specific transcriptome assemblies. Full-length transcripts were extracted from the transcriptome file using gffread and used downstream with MAKER. For *D. pseudoobscura* we used available gene expression data (RNA-seq) to create a *de novo* transcriptome as above. Gene expression data used to generate the transcriptome was from ovary, testis, male and female 3rd instar larvae, male and female whole adult, male and female carcass, male and female head, and female accessory glands (Kaiser and Bachtrog 2014). To annotate our genome assemblies, we used published protein sets from *D. pseudoobscura* and *D. melanogaster* (FlyBase) in addition to our EST libraries. The *ab initio* gene predictors SNAP (Korf 2004) and Augustus (Stanke and Waack 2003) were used to guide the annotation and we retrained the gene predictors after the first run of MAKER. In total we ran two rounds of MAKER on *D. athabasca* and *D. pseudoobscura*. For *D. subobscura* and D*. lowei*, we lacked high-quality RNA-seq data to build de novo transcriptomes and therefore we annotated these genomes using just protein sets (above) and a single round of MAKER.

Genome completeness and identification of Muller elements

To assess genome assembly completeness throughout the many assembly steps we used BUSCO (Simão *et al.* 2015) version 3.0.2 and the arthropod database (odb9). To identify the Muller elements in our assemblies and their conservation across Drosophila, we assigned 1,055 complete and single copy BUSCOs in *D. melanogaster* (dmel r6.12) to their respective Muller element and then plotted their corresponding location in each genome assembly, allowing for duplication.

Identification of obscura group orthologs

To identify orthologous genes and syntenic regions between the five genome assemblies we performed BLASTP reciprocal best hit searches between each species pair using Muller element specific proteins derived from our MAKER annotations (above). We used the blast_rbh.py script (Cock *et al.* 2015) and the gene coordinates were plotted using *genoPlotR* (Guy *et al.* 2010).

Identification of putative centromeric satellite sequence

To identify tandem repeats and putative centromeric satellites we first used Tandem Repeat Finder (TRF) (Benson 1999). We ran TRF on each genome assembly with the following settings, 2 7 7 80 10 50 500 -f -d -m. The resulting output was parsed using the k_seek.pl perl script (Wei *et al.* 2018) to collapse identical repeats and create a multi-fasta of all unique kmers in the genome. To infer the centromeric satellite

sequence we plotted the frequency of unique tandem repeat kmer lengths for each species. To identify locations of putative centromeric satellite sequences, we took all unique kmers from the inferred centromeric satellite kmer length for a given species (e.g., all related 99mers for *D. miranda*) and used Repeatmasker to mask the genome using default settings. To identify short satellite sequences we analyzed Illumina data using k-Seek (Wei *et al.* 2018) and the associated k_seek.pl script. *Drosophila guanche* Illumina data used in TRF and k-Seek analyses was downloaded from the NCBI (SRA accession, ERX2095402). To characterize tandem repeats directly from long reads we used TideHunter (Gao *et al.* 2019) with default settings and only kept consensus repeats that were > 4bp. Since the TideHunter output includes a consensus repeat sequence for each read, similar repeats and HOR variants will be found repeatedly in the output. Therefore, we clustered all consensus repeats using CD-HIT (Li and Godzik 2006) with a sequence identity cutoff of 90% and a 10bp length cutoff to help condense the output. After clustering the consensus repeats, we then BLAST searched (default settings) the representative CD-HIT cluster sequences for putative centromeric satellite sequence identified from earlier TRF analyses This allowed us to identify centromeric satellite clusters and determined their relative contribution.

Pericentromere H3K9me3 enrichment

To characterize H3K9me3 enrichment along each chromosome to help identify pericentromeric regions we generated ChIP-seq data for *D. athabasca*, *D. subobscura* and *D. lowei*, and analyzed published data for *D. pseudoobscura* (Gibilisco and Bachtrog *submitted*) and *D. miranda* (Mahajan *et al.* 2018). We generated ChIP-seq data from larval brains (*D. athabasca* EA and EB) or adult heads (*D. subobscura* and *D. lowei*) using the ultra-low-input micrococcal nuclease-based native ChIP and sequencing method (Brind'Amour *et al.* 2015) and an H3K9me3 antibody from Diagenode. *Drosophila melanogaster* was used as input for normalization. Libraries for sequencing were prepared using the SMARTer Universal Low Input DNA-seq kit (Takara/Rubicon Bio) and sequenced using 100bp PE reads on a HiSeq 4000. To quantify enrichment, raw reads from each species were mapped to their respective genome using bwa mem, downstream files processed with SAMtools, and read counts were estimated in 50kb windows using bedtools genomeCoverageBed. Counts were normalized using methods outlined in (Bonhoure *et al.* 2014). To create metagene plots we used metagene-maker (https://github.com/bdo311/metagene-maker) and normalized counts from 100bp windows.

Gene expression analyses

To characterize gene expression in *D. athabasca* we analyzed EB RNA-seq data from whole larva and pooled larval heads (above, **Supplementary File 1**). For *D. subobscura* we analyzed RNA-seq data generated from two pools of whole individuals (*Uppsalapool* and *Valenciapool*; Porcelli *et al.* 2016). For each species, all reads were aligned to their respective reference genomes using HiSat2 (Kim *et al.* 2015) and StringTie (Pertea *et al.* 2015) was used to estimate gene expression (FPKM). Downstream analyses were performed only on the subset of protein coding genes identified as orthologous between *D. subobscura* and *D. athabasca* (above).

Fluorescent in situ hybridization

DNA oligos with 5' modification or 5' and 3' modifications were used for *in situ* hybridization (**Supplementary File 6**). Tissue preparation and hybridization methods follow those of (Larracuente and Ferree 2015). Slides were imaged at the UC Berkeley Molecular Imaging Center using a Nikon Spinning Disk Confocal Microscope at 100X (with oil) with lasers 405, 488, 561, and 637 nm. Images were obtained using NIS-Elements software.

Acknowledgments

We thank Kevin Wei for discussions on satellite evolution. Thanks to Alison Nguyen, Emily Chong and Lauren Gibilisco for help in generating data. We would also like to thank J. J. Emerson for technical

assistance with PacBio data generation and Jill Bracewell for collecting *D. lowei* flies. This research was supported by NIH grants (R01GM076007, GM101255 and R01GM093182) to D. Bachtrog and by NIH grant 5F32GM123764-02 to R. Bracewell. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

REFERENCES

Allshire R. C., Karpen G. H., 2008 Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat. Rev. Genet. **9**: 923–937.

Amor D. J., Choo K. H. A., 2002 Neocentromeres: role in human disease, evolution, and centromere study. Am. J. Hum. Genet. **71**: 695–714.

Amor D. J., Bentley K., Ryan J., Perry J., Wong L., Slater H., Choo K. H. A., 2004 Human centromere repositioning "in progress". Proc. Natl. Acad. Sci. U.S.A. **101**: 6542–6547.

Ashburner M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bachmann L., Raab M., Sperlich D., 2009 Satellite DNA and speciation: A species specific satellite DNA of Drosophila guanchel. Journal of Zoological Systematics and Evolutionary Research **27**: 84–93.

Bao W., Kojima K. K., Kohany O., 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA **6**: 11.

Benson G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**: 573–580.

Bhutkar A., Schaeffer S. W., Russo S. M., Xu M., Smith T. F., Gelbart W. M., 2008 Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. Genetics **179**: 1657–1680.

Birchler J. A., Gao Z., Han F., 2009 A tale of two centromeres--diversity of structure but conservation of function in plants and animals. Funct. Integr. Genomics **9**: 7–13.

Bonhoure N., Bounova G., Bernasconi D., Praz V., Lammers F., Canella D., Willis I. M., Herr W., Hernandez N., Delorenzi M., CycliX Consortium, 2014 Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. **24**: 1157–1168.

Brind'Amour J., Liu S., Hudson M., Chen C., Karimi M. M., Lorincz M. C., 2015 An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat Commun **6**: 6033.

Brock J. A., Bloom K., 1994 A chromosome breakage assay to monitor mitotic forces in budding yeast. J. Cell. Sci. **107 (Pt 4)**: 891–902.

Brown J. D., O'Neill R. J., 2010 Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. Annu Rev Genomics Hum Genet **11**: 291–316.

Burrack L. S., Berman J., 2012 Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res. **20**: 607–619.

Burrack L. S., Hutton H. F., Matter K. J., Clancey S. A., Liachko I., Plemmons A. E., Saha A., Power E. A., Turman B., Thevandavakkam M. A., Ay F., Dunham M. J., Berman J., 2016 Neocentromeres Provide Chromosome Segregation Accuracy and Centromere Clustering to Multiple Loci along a Candida albicans Chromosome. (BG Mellone, Ed.). PLoS Genet. **12**: e1006317.

Buzzati-Traverso A. A., Scossiroli R. E., 1955 The obscura group of the genus Drosophila. Adv. Genet. **7**: 47–92.

Campbell M. S., Holt C., Moore B., Yandell M., 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics **48**: 4.11.1–39.

Carbone L., Nergadze S. G., Magnani E., Misceo D., Francesca Cardone M., Roberto R., Bertoni L., Attolini C., Francesca Piras M., de Jong P., Raudsepp T., Chowdhary B. P., Guérin G., Archidiacono N., Rocchi M., Giulotto E., 2006 Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics **87**: 777–782.

Cardone M. F., Lomiento M., Teti M. G., Misceo D., Roberto R., Capozzi O., D'Addabbo P., Ventura M., Rocchi M., Archidiacono N., 2007 Evolutionary history of chromosome 11 featuring four distinct centromere repositioning events in Catarrhini. Genomics **90**: 35–43.

Cech J. N., Peichel C. L., 2016 Centromere inactivation on a neo-Y fusion chromosome in threespine stickleback fish. Chromosome Res. **24**: 437–450.

Chang C.-H., Larracuente A. M., 2019 Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome. Genetics **211**: 333–348.

Chang C.-H., Chavan A., Palladino J., Wei X., Martins N. M. C., Santinello B., Chen C.-C., Erceg J., Beliveau B. J., Wu C.-T., Larracuente A. M., Mellone B. G., 2019 Islands of retroelements are major components of Drosophila centromeres. (PB Becker, Ed.). PLoS Biol. **17**: e3000241.

Chen C.-C., Bowers S., Lipinszki Z., Palladino J., Trusiak S., Bettini E., Rosin L., Przewloka M. R., Glover D. M., O'Neill R. J., Mellone B. G., 2015 Establishment of Centromeric Chromatin by the CENP-A Assembly Factor CAL1 Requires FACT-Mediated Transcription. Dev. Cell **34**: 73–84.

Chin C.-S., Alexander D. H., Marks P., Klammer A. A., Drake J., Heiner C., Clum A., Copeland A., Huddleston J., Eichler E. E., Turner S. W., Korlach J., 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods **10**: 563–569.

Chu C., Nielsen R., Wu Y., 2016 REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. (C Antoniewski, Ed.). PLoS ONE **11**: e0150719.

Cock P. J. A., Chilton J. M., Grüning B., Johnson J. E., Soranzo N., 2015 NCBI BLAST+ integrated into Galaxy. Gigascience **4**: 39.

Crosland M. W., Crozier R. H., 1986 Myrmecia pilosula, an Ant with Only One Pair of Chromosomes. Science **231**: 1278–1278.

DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., Hartl C., Philippakis A. A., del Angel G., Rivas M. A., Hanna M., McKenna A., Fennell T. J., Kernytsky A. M., Sivachenko A. Y., Cibulskis K., Gabriel S. B., Altshuler D., Daly M. J., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43**: 491–498.

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

Dover G., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299**: 111–117.

Dover G. A., Strachan T., Coen E. S., Brown S. D., 1982 Molecular drive. Science **218**: 1069.

Dudchenko O., Batra S. S., Omer A. D., Nyquist S. K., Hoeger M., Durand N. C., Shamim M. S., Machol I., Lander E. S., Aiden A. P., Aiden E. L., 2017 De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science **356**: 92–95.

Durand N. C., Robinson J. T., Shamim M. S., Machol I., Mesirov J. P., Lander E. S., Aiden E. L., 2016a Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst **3**: 99–101.

Durand N. C., Shamim M. S., Machol I., Rao S. S. P., Huntley M. H., Lander E. S., Aiden E. L., 2016b Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst **3**: 95–98.

Dutrillaux B., 1979 Chromosomal evolution in primates: tentative phylogeny from Microcebus murinus (Prosimian) to man. Hum. Genet. **48**: 251–314.

Elgin S. C. R., Reuter G., 2013 Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol **5**: a017780–a017780.

Federico C., Pappalardo A. M., Ferrito V., Tosi S., Saccone S., 2017 Genomic properties of chromosomal bands are linked to evolutionary rearrangements and new centromere formation in primates. Chromosome Res. **25**: 261–276.

Ferreri G. C., Liscinsky D. M., Mack J. A., Eldridge M. D. B., O'Neill R. J., 2005 Retention of latent centromeres in the Mammalian genome. J. Hered. **96**: 217–224.

Gao J.-J., Watabe H.-A., Aotsuka T., Pang J.-F., Zhang Y.-P., 2007 Molecular phylogeny of the Drosophila obscura species group, with emphasis on the Old World species. BMC Evol. Biol. **7**: 87.

Gao J.-J., Watabe H.-A., Zhang Y.-P., Aotsuka T., 1949 Karyotype Differentiation in Newly Discovered Members of the Drosophila obscura Species Group from Yunnan, China. Zoological Research **25**: 236–241.

Gao Y., Liu B., Wang Y., Xing Y., 2019 TideHunter: efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. Bioinformatics **35**: i200–i207.

Garavís M., Méndez-Lago M., Gabelica V., Whitehead S. L., González C., Villasante A., 2015 The structure of an endogenous Drosophila centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs. Sci Rep **5**: 13307.

Gel B., Serra E., 2017 karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. (J Hancock, Ed.). Bioinformatics **33**: 3088–3090.

Goubert C., Modolo L., Vieira C., ValienteMoro C., Mavingui P., Boulesteix M., 2015 De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol **7**: 1192–1205.

Gu Z., Gu L., Eils R., Schlesner M., Brors B., 2014 circlize Implements and enhances circular visualization in R. Bioinformatics **30**: 2811–2812.
Guy L., Kultima J. R., Andersson S. G. E., 2010 genoPlotR: comparative gene and genome visualization in R. Bioinformatics **26**: 2334–2335.
Hasson D., Alonso A., Cheung F., Tepperberg J. H., Papenhausen P. R., Engelen J. J. M., Warburton P. E., 2011 Formation of novel CENP-A domains on tandem repetitive DNA and across chromosome breakpoints on human chromosome 8q21 neocentromeres. Chromosoma **120**: 621–632.
Heed W. B., Crumpacker D. W., Ehrman L., 1969 Drosophila lowei, a New American Member of the Obscura Species Group1,2. Annals of the Entomological Society of America **62**: 388–393.
Henikoff S., Ahmad K., Malik H. S., 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. Science **293**: 1098–1102.
Hieter P., Pridmore D., Hegemann J. H., Thomas M., Davis R. W., Philippsen P., 1985 Functional selection and analysis of yeast centromeric DNA. Cell **42**: 913–921.
Hill T., Betancourt A. J., 2018 Extensive exchange of transposable elements in the Drosophila pseudoobscura group. Mob DNA **9**: 20.
Ishii K., Ogiyama Y., Chikashige Y., Soejima S., Masuda F., Kakuma T., Hiraoka Y., Takahashi K., 2008 Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. Science **321**: 1088–1091.
Jagannathan M., Warsinger-Pepe N., Watase G. J., Yamashita Y. M., 2017 Comparative Analysis of Satellite DNA in the Drosophila melanogaster Species Complex. G3 (Bethesda) **7**: 693–704.
Kabeche L., Nguyen H. D., Buisson R., Zou L., 2018 A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation. Science **359**: 108–114.
Kaiser V. B., Bachtrog D., 2014 De novo transcriptome assembly reveals sex-specific selection acting on evolving neo-sex chromosomes in Drosophila miranda. BMC Genomics **15**: 241.
Karageorgiou C., Gámez-Visairas V., Tarrío R., Rodríguez-Trelles F., 2019 Long-read based assembly and synteny analysis of a reference Drosophila subobscura genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. BMC Genomics **20**: 223.
Karpen G. H., Allshire R. C., 1997 The case for epigenetic effects on centromere identity and function. Trends Genet. **13**: 489–496.
Kasinathan S., Henikoff S., 2018 Non-B-Form DNA Is Enriched at Centromeres. Mol. Biol. Evol. **35**: 949–962.
Kato T., Kurahashi H., Emanuel B. S., 2012 Chromosomal translocations and palindromic AT-rich repeats. Curr. Opin. Genet. Dev. **22**: 221–228.
Khandelwal S., 2008 Chromosome evolution in the genus Ophioglossum L. Botanical Journal of the Linnean Society **102**: 205–217.
Khost D. E., Eickbush D. G., Larracuente A. M., 2017 Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. Genome Res. **27**: 709–721.
Kim D., Langmead B., Salzberg S. L., 2015 HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**: 357–360.
Koren S., Walenz B. P., Berlin K., Miller J. R., Bergman N. H., Phillippy A. M., 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**: 722–736.
Korf I., 2004 Gene finding in novel genomes. BMC Bioinformatics **5**: 59.
Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., Antonescu C., Salzberg S. L., 2004 Versatile and open software for comparing large genomes. Genome Biol. **5**: R12.
Larracuente A. M., Ferree P. M., 2015 Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. J Vis Exp: 52288.
Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences. (I Birol, Ed.). Bioinformatics **34**: 3094–3100.
Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760.
Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.
Li W., Godzik A., 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**: 1658–1659.
Lieberman-Aiden E., van Berkum N. L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B. R., Sabo P.

J., Dorschner M. O., Sandstrom R., Bernstein B., Bender M. A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L. A., Lander E. S., Dekker J., 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**: 289–293.

Liehr T., Kosyakova N., Weise A., Ziegler M., Raabe-Meyer G., 2010 First case of a neocentromere formation in an otherwise normal chromosome 7. Cytogenet. Genome Res. **128**: 189–191.

Lohe A. R., Hilliker A. J., Roberts P. A., 1993 Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. Genetics **134**: 1149–1174.

Lomiento M., Jiang Z., D'Addabbo P., Eichler E. E., Rocchi M., 2008 Evolutionary-new centromeres preferentially emerge within gene deserts. Genome Biol. **9**: R173.

Maggert K. A., Karpen G. H., 2001 The activation of a neocentromere in Drosophila requires proximity to an endogenous centromere. Genetics **158**: 1615–1628.

Mahajan S., Wei K. H.-C., Nalley M. J., Gibilisco L., Bachtrog D., 2018 De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. (C Tyler-Smith, Ed.). PLoS Biol. **16**: e2006348.

Malik H. S., Henikoff S., 2009 Major evolutionary transitions in centromere complexity. Cell **138**: 1067–1082.

Marshall O. J., Chueh A. C., Wong L. H., Choo K. H. A., 2008 Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am. J. Hum. Genet. **82**: 261–282.

McGurk M. P., Barbash D. A., 2018 Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. Genome Res. **28**: 714–725.

McNulty S. M., Sullivan L. L., Sullivan B. A., 2017 Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C. Dev. Cell **42**: 226–240.e6.

Melters D. P., Bradnam K. R., Young H. A., Telis N., May M. R., Ruby J. G., Sebra R., Peluso P., Eid J., Rank D., Garcia J. F., DeRisi J. L., Smith T., Tobias C., Ross-Ibarra J., Korf I., Chan S. W. L., 2013 Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. **14**: R10.

Miller D. E., Staber C., Zeitlinger J., Hawley R. S., 2018 Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 (Bethesda) **8**: 3131–3141.

Muller H. J., 1940 Bearings of the Drosophila work on systematics. In:JS H (Ed.), *The new Systematics*, Oxford.

Nishimura K., Komiya M., Hori T., Itoh T., Fukagawa T., 2019 3D genomic architecture reveals that neocentromeres associate with heterochromatin regions. J. Cell Biol. **218**: 134–149.

O'Neill R. J., Eldridge M. D. B., Metcalfe C. J., 2004 Centromere dynamics and chromosome evolution in marsupials. J. Hered. **95**: 375–381.

Orengo D. J., Puerma E., Aguade M., 2019 The molecular characterization of fixed inversions breakpoints unveils the ancestral character of the Drosophila guanche chromosomal arrangements. Sci Rep **9**: 1706.

Patterson J. T., Stone W. S., 1952 *Evolution in the genus Drosophila*. The Macmillan Company, New York.

Pertea M., Pertea G. M., Antonescu C. M., Chang T.-C., Mendell J. T., Salzberg S. L., 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**: 290–295.

Porcelli D., Westram A. M., Pascual M., Gaston K. J., Butlin R. K., Snook R. R., 2016 Gene expression clines reveal local adaptation and associated trade-offs at a continental scale. Sci Rep **6**: 32975.

Powell J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York, Oxford.

Puerma E., Orengo D. J., Cruz F., Gómez-Garrido J., Librado P., Salguero D., Papaceit M., Gut M., Segarra C., Alioto T. S., Aguade M., 2018 The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability. (J Gonzalez, Ed.). Genome Biol Evol **10**: 1956–1969.

Quinlan A. R., Hall I. M., 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842.

Ramani V., Cusanovich D. A., Hause R. J., Ma W., Qiu R., Deng X., Blau C. A., Disteche C. M., Noble W. S., Shendure J., Duan Z., 2016 Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc **11**: 2104–2121.

Rhoades M. M., Dempsey E., 1966 The Effect of Abnormal Chromosome 10 on Preferential Segregation and Crossing over in Maize. Genetics **53**: 989–1020.

Riddle N. C., Jung Y. L., Gu T., Alekseyenko A. A., Asker D., Gui H., Kharchenko P. V., Minoda A., Plachetka A., Schwartz Y. B., Tolstorukov M. Y., Kuroda M. I., Pirrotta V., Karpen G. H., Park P. J., Elgin S. C. R., 2012 Enrichment of HP1a on Drosophila chromosome 4 genes creates an alternate chromatin structure critical for

regulation in this heterochromatic domain. (JD Lieb, Ed.). PLoS Genet. **8**: e1002954.
Riddle N. C., Minoda A., Kharchenko P. V., Alekseyenko A. A., Schwartz Y. B., Tolstorukov M. Y., Gorchakov A. A., Jaffe J. D., Kennedy C., Linder-Basso D., Peach S. E., Shanower G., Zheng H., Kuroda M. I., Pirrotta V., Park P. J., Elgin S. C. R., Karpen G. H., 2011 Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. **21**: 147–163.
Rocchi M., Archidiacono N., Schempp W., Capozzi O., Stanyon R., 2012 Centromere repositioning in mammals. Heredity (Edinb) **108**: 59–67.
Roizès G., 2006 Human centromeric alphoid domains are periodically homogenized so that they vary substantially between homologues. Mechanism and implications for centromere functioning. Nucleic Acids Res. **34**: 1912–1924.
Sart du D., Cancilla M. R., Earle E., Mao J. I., Saffery R., Tainton K. M., Kalitsis P., Martyn J., Barry A. E., Choo K. H., 1997 A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. Nat. Genet. **16**: 144–153.
Sato H., Masuda F., Takayama Y., Takahashi K., Saitoh S., 2012 Epigenetic inactivation and subsequent heterochromatinization of a centromere stabilize dicentric chromosomes. Curr. Biol. **22**: 658–667.
Schaeffer S. W., Bhutkar A., McAllister B. F., Matsuda M., Matzkin L. M., O'Grady P. M., Rohde C., Valente V. L. S., Aguade M., Anderson W. W., Edwards K., Garcia A. C. L., Goodman J., Hartigan J., Kataoka E., Lapoint R. T., Lozovsky E. R., Machado C. A., Noor M. A. F., Papaceit M., Reed L. K., Richards S., Rieger T. T., Russo S. M., Sato H., Segarra C., Smith D. R., Smith T. F., Strelets V., Tobari Y. N., Tomimura Y., Wasserman M., Watts T., Wilson R., Yoshida K., Markow T. A., Gelbart W. M., Kaufman T. C., 2008 Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics **179**: 1601–1655.
Schneider K. L., Xie Z., Wolfgruber T. K., Presting G. G., 2016 Inbreeding drives maize centromere evolution. Proc. Natl. Acad. Sci. U.S.A. **113**: E987–96.
Schubert I., 2018 What is behind "centromere repositioning"? Chromosoma **127**: 229–234.
Scott K. C., Sullivan B. A., 2014 Neocentromeres: a place for everything and everything in its place. Trends Genet. **30**: 66–74.
Segarra C., Aguadé M., 1992 Molecular organization of the X chromosome in different species of the obscura group of Drosophila. Genetics **130**: 513–521.
Segarra C., Lozovskaya E. R., Ribó G., Aguadé M., Hartl D. L., 1995 P1 clones from Drosophila melanogaster as markers to study the chromosomal evolution of Muller's A element in two species of the obscura group of Drosophila. Chromosoma **104**: 129–136.
Shang W.-H., Hori T., Martins N. M. C., Toyoda A., Misu S., Monma N., Hiratani I., Maeshima K., Ikeo K., Fujiyama A., Kimura H., Earnshaw W. C., Fukagawa T., 2013 Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. Dev. Cell **24**: 635–648.
She X., Horvath J. E., Jiang Z., Liu G., Furey T. S., Christ L., Clark R., Graves T., Gulden C. L., Alkan C., Bailey J. A., Sahinalp C., Rocchi M., Haussler D., Wilson R. K., Miller W., Schwartz S., Eichler E. E., 2004 The structure and evolution of centromeric transition regions within the human genome. Nature **430**: 857–864.
Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M., 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**: 3210–3212.
Smith A., Hubley R., Green P., *RepeatMasker Open-4.0*. httpwww.repeatmasker.org.
Stanke M., Waack S., 2003 Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19 Suppl 2**: ii215–25.
Steiner F. A., Henikoff S., 2015 Diversity in the organization of centromeric chromatin. Curr. Opin. Genet. Dev. **31**: 28–35.
Stephan W., Cho S., 1994 Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics **136**: 333–341.
Sturtevant A. H. A., 1936 Observations on the Species Related to New Forms of Drosophila affinis, with Descriptions of Seven. The American Naturalist **70**: 574–584.
Sun X., Wahlstrom J., Karpen G., 1997 Molecular structure of a functional Drosophila centromere. Cell **91**: 1007–1019.
Talbert P. B., Henikoff S., 2018 Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly. Trends Genet. **34**: 587–599.
Talbert P., Kasinathan S., Henikoff S., 2018 Simple and Complex Centromeric Satellites in Drosophila Sibling

Species. Genetics: genetics.300620.2017.
Tolomeo D., Capozzi O., Stanyon R. R., Archidiacono N., D'Addabbo P., Catacchio C. R., Purgato S., Perini G., Schempp W., Huddleston J., Malig M., Eichler E. E., Rocchi M., 2017 Epigenetic origin of evolutionary novel centromeres. Sci Rep **7**: 41980.
Török T., Gorjánácz M., Bryant P. J., Kiss I., 2000 Prod is a novel DNA-binding protein that binds to the 1.686 g/cm(3) 10 bp satellite repeat of Drosophila melanogaster. Nucleic Acids Res. **28**: 3551–3557.
Tree of Sex Consortium, 2014 Tree of Sex: a database of sexual systems. Sci Data **1**: 140015.
van Hooff J. J., Tromer E., van Wijk L. M., Snel B., Kops G. J., 2017 Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep. **18**: 1559–1571.
Vaser R., Sović I., Nagarajan N., Šikić M., 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. **27**: 737–746.
Ventura M., Antonacci F., Cardone M. F., Stanyon R., D'Addabbo P., Cellamare A., Sprague L. J., Eichler E. E., Archidiacono N., Rocchi M., 2007 Evolutionary formation of new centromeres in macaque. Science **316**: 243–246.
Walker B. J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C. A., Zeng Q., Wortman J., Young S. K., Earl A. M., 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. (J Wang, Ed.). PLoS ONE **9**: e112963.
Wang G., Zhang X., Jin W., 2009 An overview of plant centromeres. J Genet Genomics **36**: 529–537.
Wei K. H.-C., Grenier J. K., Barbash D. A., Clark A. G., 2014 Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. Proc. Natl. Acad. Sci. U.S.A. **111**: 18793–18798.
Wei K. H.-C., Lower S. E., Caldas I. V., Sless T. J. S., Barbash D. A., Clark A. G., 2018 Variable Rates of Simple Satellite Gains across the Drosophila Phylogeny. Mol. Biol. Evol. **35**: 925–941.
Whiting J. H., Plliley M. D., Farmer J. L., Jeffery D. E., 1989 In situ hybridization analysis of chromosomal homologies in Drosophila melanogaster and Drosophila virilis. Genetics **122**: 99–109.
Wick R. R., Schultz M. B., Zobel J., Holt K. E., 2015 Bandage: interactive visualization of de novo genome assemblies. Bioinformatics **31**: 3350–3352.
Willard H. F., 1990 Centromeres of mammalian chromosomes. Trends Genet. **6**: 410–416.
Wong Miller K. M., Bracewell R. R., Eisen M. B., Bachtrog D., 2017 Patterns of Genome-Wide Diversity and Population Structure in the Drosophila athabasca Species Complex. Mol. Biol. Evol. **34**: 1912–1923.
Zeitlin S. G., Baker N. M., Chapados B. R., Soutoglou E., Wang J. Y. J., Berns M. W., Cleveland D. W., 2009 Double-strand DNA breaks recruit the centromeric histone CENP-A. Proc. Natl. Acad. Sci. U.S.A. **106**: 15762–15767.
Zhang H., Freudenreich C. H., 2007 An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. Mol. Cell **27**: 367–379.

FIGURE AND TABLE CAPTIONS

Figure 1. Phylogenetic relationships and karyotype evolution in the *D. obscura* group. *Drosophila subobscura* represents the ancestral karyotype condition consisting of five large and one small pair of telocentric chromosomes (termed Muller elements A-F). Phylogeny adapted from (Gao *et al.* 2007). Chromosomal fusions and movement of centromeres along the chromosomes has resulted in different karyotypes in different species groups (Segarra *et al.* 1995; Schaeffer *et al.* 2008). Indicated along the tree are transitions of chromosome morphology, and the different subgroups of the *obscura* species group are indicated by gray shading (the *subobscura*, *obscura*, *pseudoobscura*, and *affinis* subgroup). Muller elements are color coded, and centromeres are shown as black circles.

Figure 2. Genome organization in *Drosophila obscura* group flies. Shown here are the assembled chromosome sizes, scaffolding stitch points, gene density, repeat content (percentage of bases repeat-masked in 100kb windows) and H3K9me3 enrichment (50kb windows) across the genome assemblies of *D. subobscura*, *D. athabasca*, *D. lowei*, *D. pseudoobscura* and *D. miranda*. Muller elements are color coded,

gene density is shown as a black to green heatmap (genes per 100kb) , H3K9me3 enrichment is shown in orange, and repeat density is shown in teal (note that H3K9me3 enrichment and repeat density are plotted semi-transparent). Scaffolding stich points are indicated as vertical lines.

Figure 3. Chromosome synteny and evolution. A. Conservation of Muller elements in the Drosophila genus. Orthologous single copy *Drosophila melanogaster* (Dmel) BUSCOs plotted on reference genome assemblies. Muller elements are color-coded based on *D. melanogaster*. **B.** Comparisons of synteny between our genome assemblies. Muller elements are color-coded based on the *D. subobscura* genome. Each line represents a protein-coding gene. Ovals denote the location of the putative centromere (based on the location of centromere-associated satellite sequences, see **Fig. 4**).

Figure 4. Identification of centromere-associated satellite sequences. **A.** Histograms of most abundant satellites in assembled genomes. Repeat length refer to the size of the repeat unit. For each species apart from *D. subobscura*, a specific satellite (or higher-order variant of it as indicated by the same colors) is enriched. In *D. miranda*, a 99mer (in green) and four units of a unrelated 21mer (84bp; in red) are the most abundant satellites, in *D. pseudoobscura*, four units of a similar 21mer (84bp; in red) is most common, in *D. lowei*, three units of a similar 21mer (63bp; in red) is most common, in *D. athabasca*, an unrelated 160mer (in pink) is the most common satellite, and in the *D. guanche* genome (Puerma *et al.* 2018), an unrelated 290mer (in black) is most common. No abundant satellite was identified in the assembled genome of *D. subobscura*. **B**. Location of putative centromere-associated repeats (from panel A) in pericentromeric regions. In *D. suboobscura* a 12mer is highly enriched in raw sequencing reads. Shown is a 5 Mb fragment for each chromosome with the highest density of the satellite sequence (that is the putative centromere), and all unplaced scaffolds. **C.** FISH hybridization confirms centromere location of identified satellites (same color coding as in A and B). Probes corresponding to the 21mer (Cy5; red) and 99mer (Cy3; green) were hybridized to both *D. miranda* and *D. pseudoobscura*; the 21mer showed a centromere location in both species, while only the 99mer hybridized to the centromeres of *D. miranda*. The 160mer (6FAM; red) localized to the centromeres of *D. athabasca* chromosome, and the 12mer (TYE665; red) to the centromeres of *D. subobscura* chromosomes. Stronger hybridization signal supposedly correspond to higher repeat abundance at a particular genomic location.

Figure 5. Emergence and loss of centromeres. A. Shown are homologous genes between *D. subobscura* (telocentric), *D. athabasca* (metacentric) and *D. pseudoobscura* (metacentric and telocentric) with H3K9me3 enrichment plotted along Muller A (red), B (green) and E (purple) in 50kb windows. Genes identified in the pericentromere of metacentric chromosomes are shown with black lines. Genes identified in pericentromeres of metacentric chromosomes can be traced to two 'seed regions' each on the telocentric chromosome of *D. suboboscura*, and to paleocentromere regions in species that secondarily lost the metacentric centromere. **B.** GC-content across *D. subobscura* Muller A, B and E. Seed regions have significantly lower GC-content compared to genomic background (**Supplementary File 7**).

Figure 6. Karyotype and centromere evolution. A. Models for transitions between metacentric and telocentric chromosomes, either invoking pericentric inversions (top), or centromere repositioning (bottom) via the birth of a new centromere (lightning bolt) and death of the old centromere (skull and crossbones). The pericentromere is indicated by darker shading, the centromere as a white rectangle. **B.** The syntenic location of genes adjacent to the centromere can allow us to distinguish between a simple inversion models vs. centromere relocation. The genes closest to the centromere of the telocentric chromosome (30 genes in panel C) are shown by different shading. **C.** Dot plots for homologous genes (semi-transparent points) between telocentric and metacentric Muller elements (orange: Muller A-AD; purple: Muller E; green: Muller B). In 4 out of 5 cases, pericentric genes in the telocentric species are found in the non-pericentric regions of the metacentric species. Only Muller B between *D. athabasca* and

pseudoobscura group flies (*D. lowei* is pictured) shows that the same genes are pericentric in both species (and thus support a simple inversion model).

Figure 7. Functional consequences of becoming pericentromeric. A. Metagene plots showing H3K9me3 enrichment for genes located in different parts of the genome in *D. subobscura* (top) and *D. athabasca* (bottom). **B**. Patterns of gene expression for homologous genes in *D. subobscura* and *D. athabasca,* classified as whether they are part of the 'seed' region in *D. subobscura* that become part of the pericentromeric heterochromatin in *D. athabasca* or not. Expression patterns were not found to significantly differ between *D. subobscura* non-pericentromeric genes and seed genes, while seed orthologs located in the pericentromere of *D. athabasca* showed significantly higher expression than non-pericentromeric genes (Mann-Whitney U, $p < 0.0001$).

Figure 8. Transposable element evolution across the genome. Shown is the fraction of bases masked in a 100kb genomic windows for different transposable element families with the total TE fraction plotted above each chromosome.

SUPPLEMENTAL FIGURE AND TABLE CAPTIONS

Figure 2 - figure supplement 1. Illumina coverage over assembled chromosomes in reference genome assemblies of *D. subobscura*, *D pseudoobscura, D. lowei and D. athabasca*. Male (blue) and female (red) Illumina coverage (log2) shown in 5kb non-overlapping windows along each chromosome. Scaffolding stitch point locations shown in the outermost track. Note that stitch points often coincide with aberrant patterns in Illumina coverage highlighting difficult regions to assemble.

Figure 2 - figure supplement 2. Nanopore or PacBio coverage over reference genome assemblies of *D. subobscura*, *D pseudoobscura, D. lowei and D. athabasca*. Coverage (log2) shown in 50kb non-overlapping windows along the genome assembly with different genomic partitions highlighted. Contigs < 50kb not shown. Note, *D. lowei* data was derived from males, *D. subobscura* and *D. athabasca* was derived from females, and *D. pseudoobscura* was derived from a combination of males and females. Therefore, coverage over A-AD is expected to be lower in *D. lowei* and *D. pseudoobscura* relative to the autosomes (Mullers B, C, E, F). Further, in *D. pseudoobscura*, A-AD and Y coverage is not expected to be similar due to differences in the female/male contribution to the total read pool.

Figure 2 - figure supplement 3. Hi-C association heatmaps of genome assemblies for *D. subobscura*, *D pseudoobscura, D. lowei and D. athabasca*. Lines demarcate assembled chromosomes and unplaced or Y contigs. Note that pericentromeric regions typically show weak Hi-C associations with neighboring euchromatic regions due to their repetitiveness, thus producing 'checkerboard' patterns in the plot. Also of note, off-diagonal associations are apparent in many euchromatic arms in our hybrid *D. athabasca* assembly and thus identify chromosomal inversions (see methods).

Figure 3 - figure supplement 1. Whole genome alignments (MUMmer) of our *Drosophila subobscura* genome assembly (strain 14011-0131.10) with A) *D. subobscura* strain *ch-cu* and B) *D. guanche*. We show the Muller element naming scheme on the X axis and the *subobscura* group chromosome naming scheme on the Y axis.

Figure 3 - figure supplement 2. Whole genome alignments (MUMmer) of draft *Drosophila bifasciata* genome contigs with A) both Muller A and D of *D. subobscura* and B) highlighting the Muller A centromere seed regions. Many small contigs in the *D. bifasciata* assembly map to the *D. subobscura* seed region which is indicative of a highly repetitive and more poorly assembled region (i.e., pericentromere) in *D. bifasciata*.

C) Alignments of *D. bifasciata* contigs with the fused Muller A-AD of *D. pseudoobscura* show sequence similarity over the large metacentric pericentromere of Muller A-AD.

Figure 4 - figure supplement 1. Genomic distribution of inferred centromeric satellite sequences.

Figure 4 - figure supplement 2. Short satellite DNAs in obscura group flies. Shown is a heatmap of the results from k-Sseek analyses used to identify enriched satellite sequences. Only kmers > 1bp that constitute > 10% of the total short satellite sequence in any one species are shown. White = not present, red = highly enriched.

Figure 4 - figure supplement 3. Identification of centromere-associated satellite sequences from Nanopore and PacBio reads. Shown are counts of satellite lengths identified directly from raw sequencing reads for each Drosophila species. Colored lines are drawn as in Fig. 4 to highlight overlap between the results from TRF analyses and TideHunter analyses.

Figure 4 - figure supplement 4. Additional fluorescent *in situ* hybridization images. A) To help visualize enrichment on all chromosomes in species with variable intensities, A) shows only the color channel that identifies the 160mer and 12mer in *D. athabasca* and *D. subobscura,* respectively. Arrows show low intensity signal on Muller C in *D. athabasca* and an unknown Muller element in *D. subobscura*. B) Shows replicates for each species and satellite placement during cell division.

Figure 4 - figure supplement 5. Nanopore and Illumina sequencing coverage over unplaced contigs (Contig_6 and Contig_7) harboring arrays of putative *D. subobscura* centromeric-associated satellite sequence.

Figure 5 - figure supplement 1. Alignments of Muller A between *D. subobscura* (telocentric) with *D. pseudoobscura* (metacentric) and H3K9me3 enrichment plotted in 50kb windows above each chromosome. Pericentromeric genes in *D. pseudoobscura* shown in black.

Figure 5 - figure supplement 2. GC-content, the percentage of bases repeat-masked, and number of genes, in 10kb non-overlapping windows across Muller A, B and E. Seed regions are shown bounded by dashed lines. Note the orientation for the Muller elements are shown as in Fig. 5B.

Figure 5 - figure supplement 3. GC-content of different functional categories in seed and non-seed regions of Muller A, B and E of *D. subobscura*.

Figure 8 - figure supplement 1. Genomic distribution of transposable elements by species and Muller element. The top 10 TE's per Muller element, per species, are shown in descending order from top to bottom and ranked by their total contribution (bp) to each element. Each point represents a genomic location masked for an element.

Figure 8 - figure supplement 2. De-novo estimates (dnaPipeTE) of transposable element frequencies in *D. subobscura*, *D. athabasca*, *D. lowei* and *D. pseudoobscura*.

Supplementary File 1. DNA sequence data generated for this study.

Supplementary File 2. Drosophila strains used for genome assembly.

Supplementary File 3. BUSCO results for assembled genomes. For *D. athabasca*, BUSCO scores for the EB assembly are shown.

Supplementary File 4. Number of protein coding gene models from MAKER annotations for each genome assembly and Muller element.

Supplementary File 5. Average percentage of bases repeat-masked in each pericentromere.

Supplementary File 6. Inferred centromeric satellite sequence and fluorescent *in situ* hybridization probes.

Supplementary File 7. Comparison of gene density, repeat density, and GC-content (%), between seed and non-seed regions of Muller A, B and E in *D. subobscura*. *P*-values for comparisons, Mann-Whitney U.

Supplementary File 8. The most common repeat families in each species and amount of masked sequence (bp).

Supplementary File 9. The top 10 TE's from each Muller element and amount of masked sequence (bp).

Supplementary File 10. Hi-C data summary.

Appendix 1. Details of assembly procedures for each species.

Table 1. Total length (bp) of each assembled Muller element in each species, the number of contigs, and estimated length (Mb) of pericentromere sequence.

Species		Muller A	Muller D	Muller A-AD	Muller B	Muller C	Muller E	Muller F	total
D. subobscura									
	chromosome (bp)	24,182,865	23,815,339	n/a	25,941,769.00	20,343,353.00	30,159,154.00	1,505,893.00	125,948,373.00
	contigs	9	8	n/a	5	1	3	4	30
	pericentromere (Mb)	1.9	1.7	n/a	2.8	1.1	1.1	n/a	8.6
D. athabasca									
	chromosome (bp)	n/a	n/a	67,112,822.00	52,101,127.00	24,053,775.00	42,973,490.00	1,524,173.00	187,765,387.00
	contigs	n/a	n/a	4	4	6	7	1	22
	pericentromere (Mb)	n/a	n/a	14.1	22.8	2.5	11.2	n/a	50.6
D. lowei									
	chromosome (bp)	n/a	n/a	73,251,623.00	31,032,897.00	24,430,087.00	48,132,706.00	1,606,711.00	178,454,024.00
	contigs	n/a	n/a	190	42	47	152	1	432
	pericentromere (Mb)	n/a	n/a	17.2	2.2	3.5	15.1	n/a	38
D. miranda									
	chromosome (bp)	n/a	n/a	77,621,844.00	32,539,841.00	25,306,191.00	35,263,383.00	2,366,016.00	173,097,275.00
	contigs	n/a	n/a	18	2	3	3	1	27
	pericentromere (Mb)	n/a	n/a	20.5	3.4	3.4	2	n/a	29.3
D. pseudoobscura									
	chromosome (bp)	n/a	n/a	67,434,674.00	30,637,803.00	22,641,560.00	32,023,297.00	1,941,385.00	154,678,719.00
	contigs	n/a	n/a	37	6	5	5	1	54
	pericentromere (Mb)	n/a	n/a	14.1	2.8	2.7	0.7	n/a	20.3

Table 2. Putative centromeric satellite lengths inferred from Tandem Repeat Finder (Benson 1999), k-Seek (Wei *et al.* 2018), and TideHunter (Gao *et al.* 2019) for each Drosophila species. HOR = higher order repeat.

	Tandem Repeat Finder	***k-Seek***	***TideHunter***
D. subobscura	no candidate	12-bp	12-bp, 107-bp
D. athabasca	160-bp	11-bp	11-bp, 160-bp
D. lowei	63-bp(HOR of 21-bp)	no candidate	21-bp
D. miranda	99-bp, 84-bp(HOR of 21-bp)	no candidate	99-bp, 84-bp(HOR of 21-bp)
D. pseudoobscura	84-bp(HOR of 21-bp)	no candidate	168-bp(HOR of 21-bp)

Table 3. Transposable elements by Muller element.

Species	TE	total bp masked	% of genome masked
D. subobscura			
	total TE's	7,572,806	6.0%
	Dpse_Gypsy_6	1,319,782	1.0%
	CR1-1_DPer	449,484	0.4%
	Gypsy8-I_Dpse	424,330	0.3%
	LOA-1_DPer	331,166	0.3%
	T213_X.Unknown	323,480	0.3%
D. athabasca			
	total TE's	42,382,296	22.6%
	Daff_Jockey_18	3,779,133	2.0%
	CR1-1_DPer	2,938,986	1.6%
	T32_LTR	1,958,710	1.0%
	LOA-1_DPer	1,729,016	0.9%
	LOA-2_DPer	1,683,407	0.9%
D. lowei			
	total TE's	45,307,006	25.4%
	HelitronN-1_DPe	2,764,830	1.5%
	CR1-1_DPer	2,564,981	1.4%
	LOA-3_DPer	1,421,120	0.8%
	LOA-2_DPer	1,019,792	0.6%
	BEL-3_DPer-I	965,011	0.5%
D. pseudoobscura			
	total TE's	29,907,407	19.3%
	CR1-1_DPer	2,082,339	1.3%
	HelitronN-1_DPe	1,962,822	1.3%
	LOA-3_DPer	1,024,994	0.7%
	T154_X.Unknown	799,007	0.5%
	LOA-2_DPer	726,556	0.5%
D. miranda			
	total TE's	42,680,234	24.7%
	HelitronN-1_DPe	3,834,913	2.2%
	CR1-1_DPer	3,148,617	1.8%
	Gypsy18-I_Dpse	2,261,130	1.3%
	LOA-3_DPer	1,557,989	0.9%
	LOA-2_DPer	1,208,941	0.7%

D. athabasca

D. affinis

affinis

D. lowei

A, D fusion

B telocentric

D. pseudoobscura

D. persimilis

pseudoobscura

A, B, E
metacentric

E telocentric

D. miranda

D. bifasciata

obscura

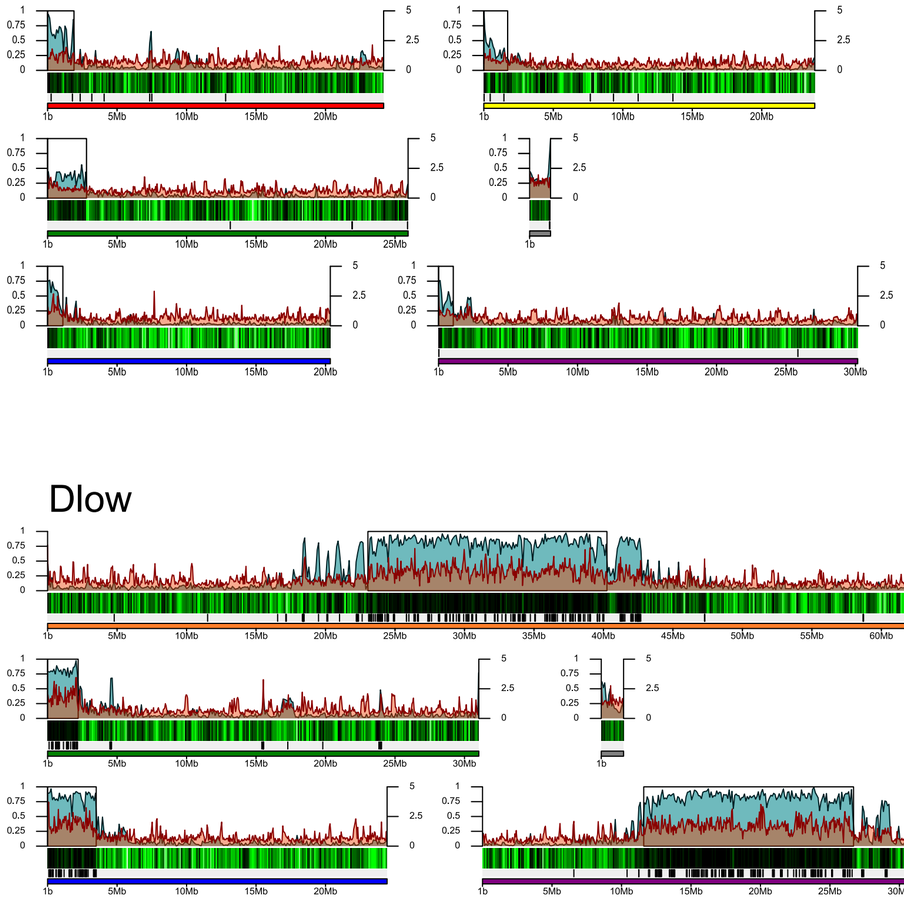
D. guanche

subobscura

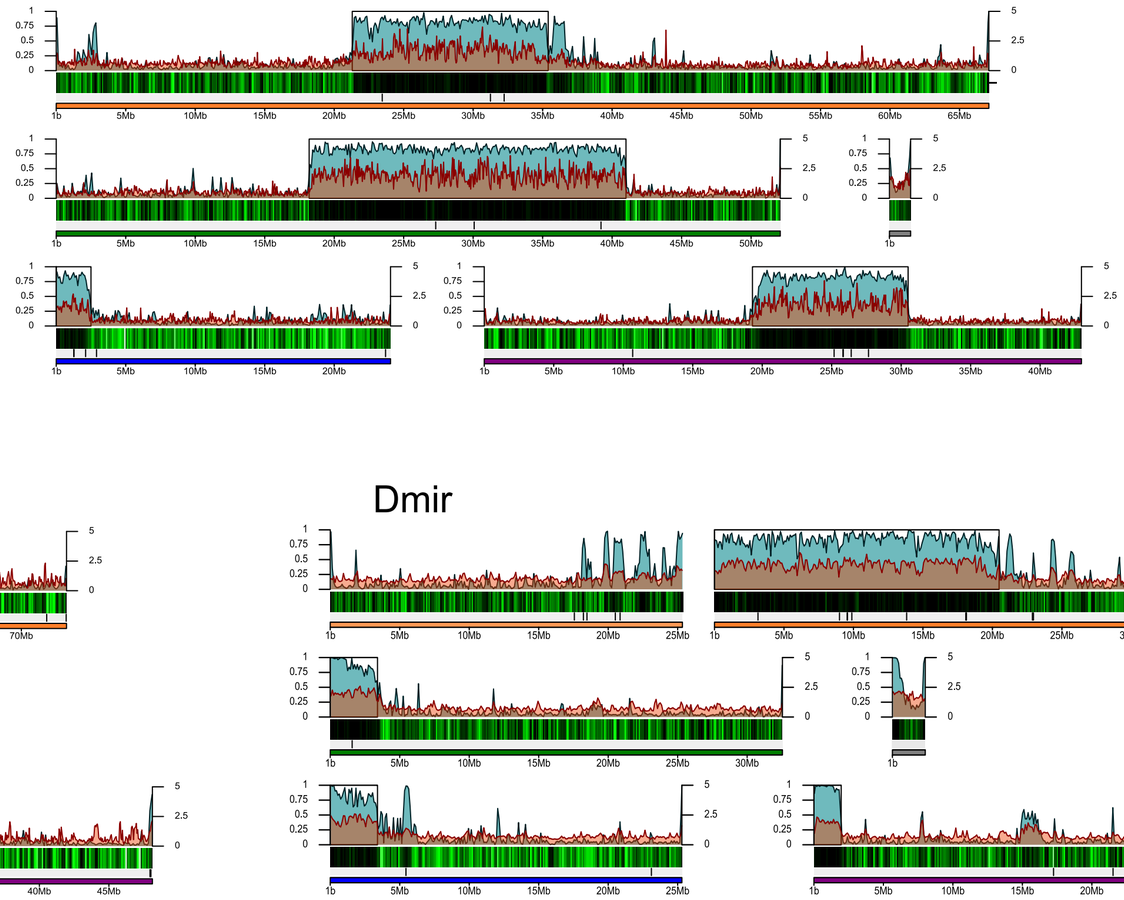
D. subobscura

Muller A
Muller B
Muller C
Muller D
Muller E
Muller F
Centromere

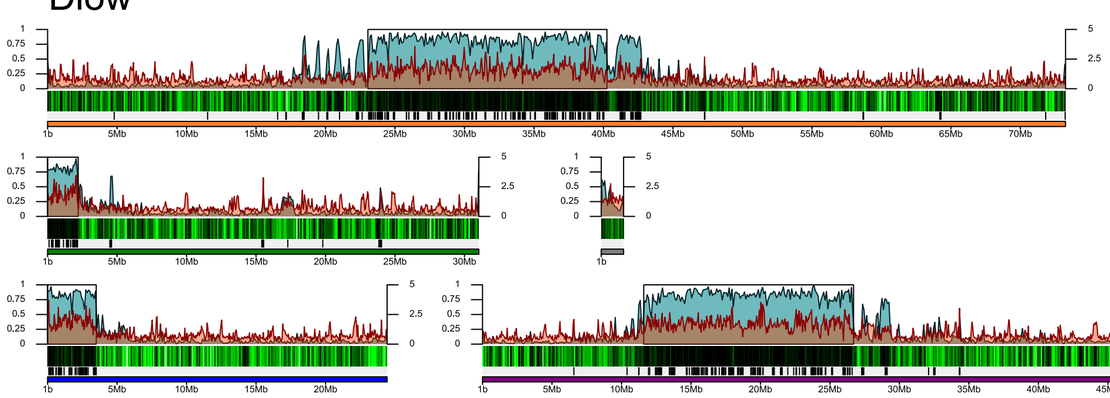
Dsub



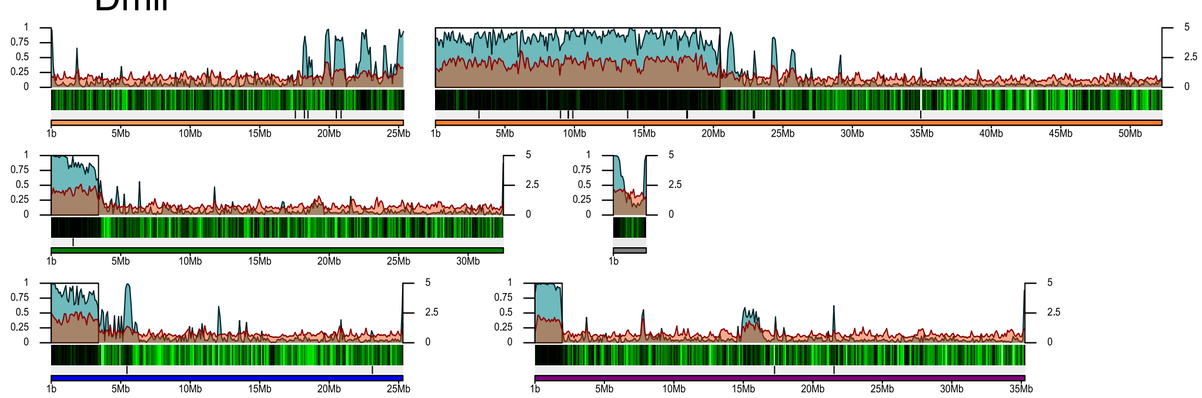
Dath



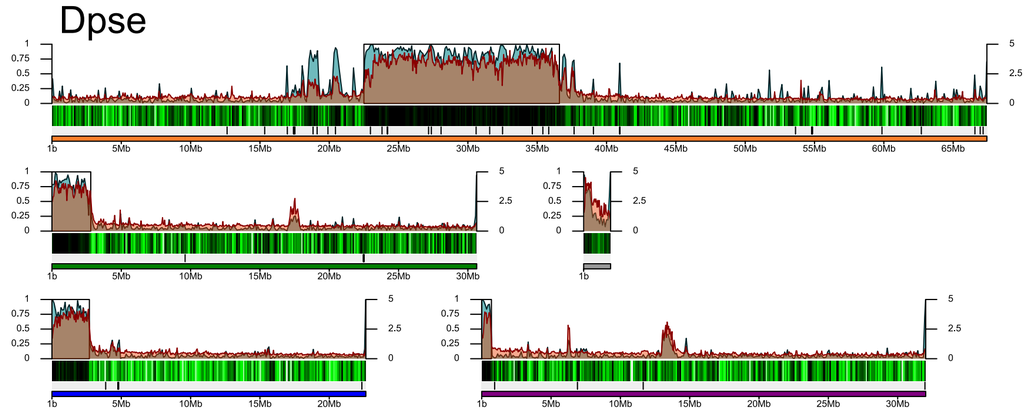
Dlow



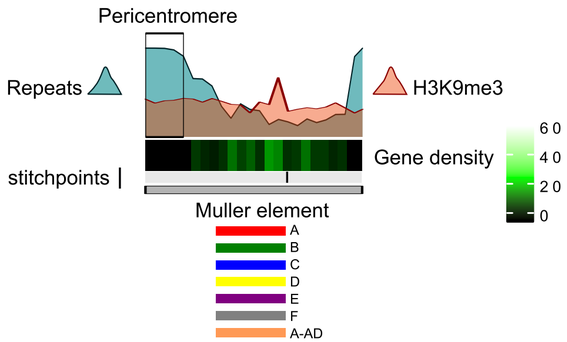
Dmir



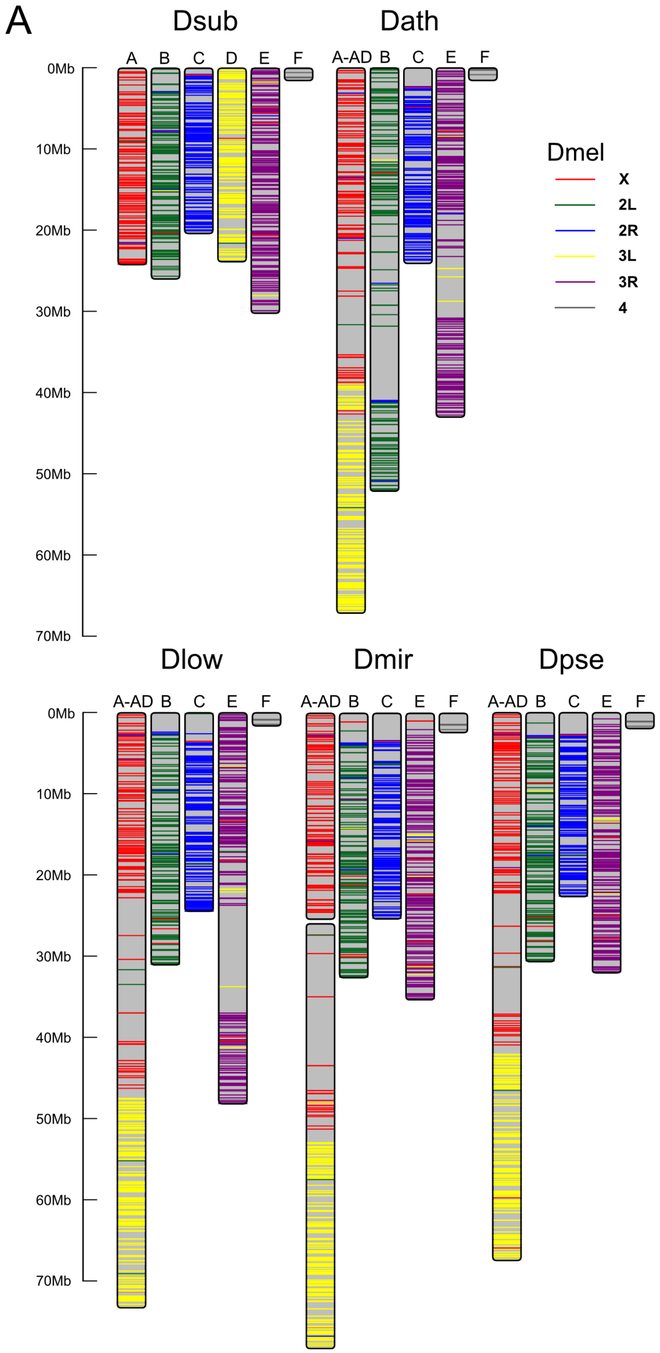
Dpse



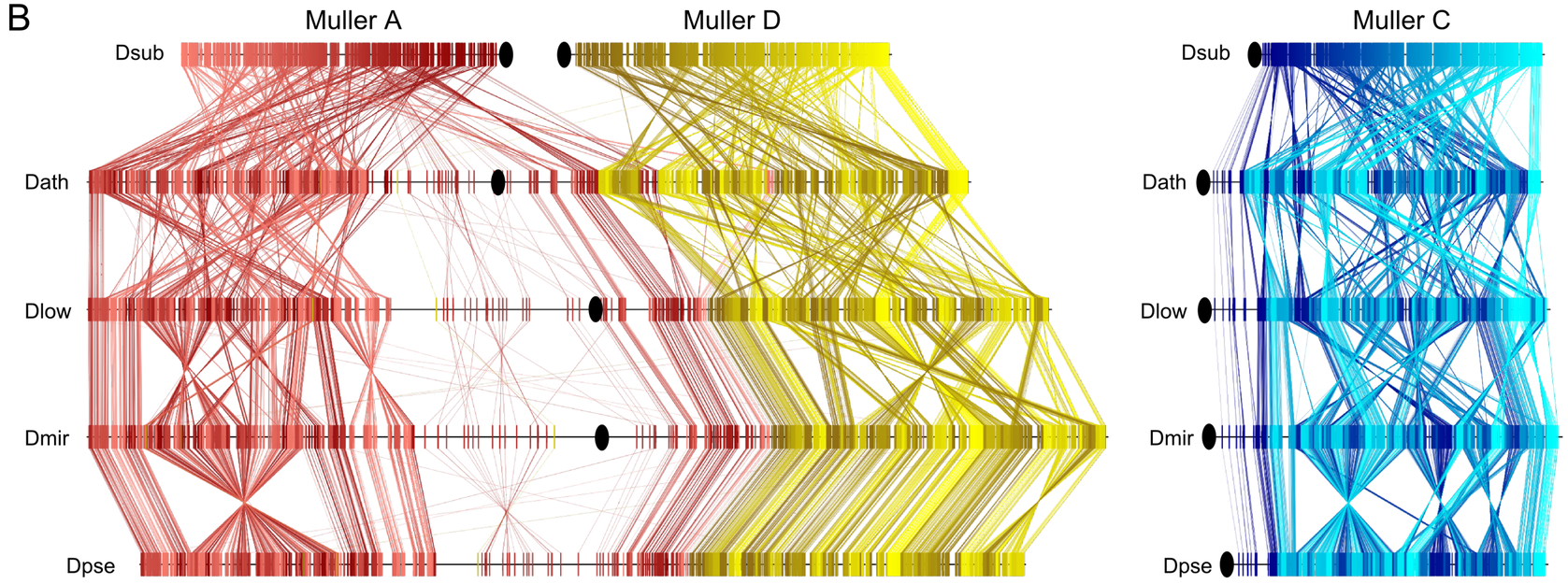
Legend

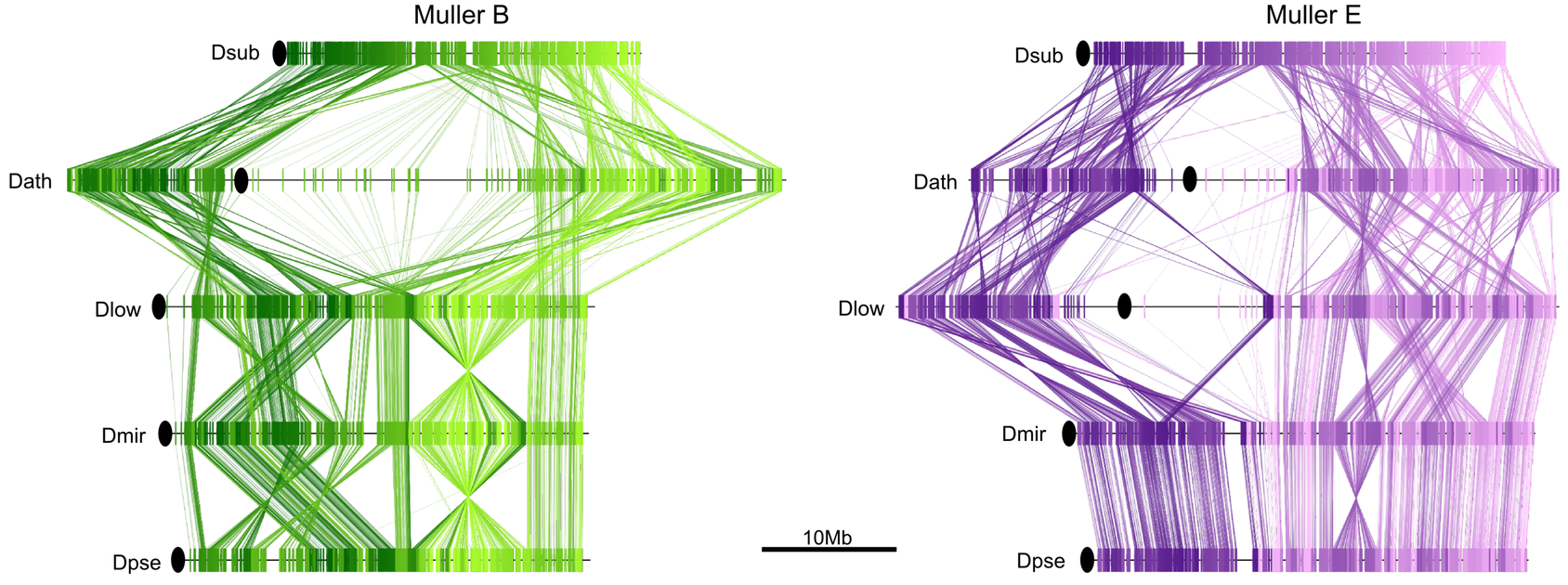


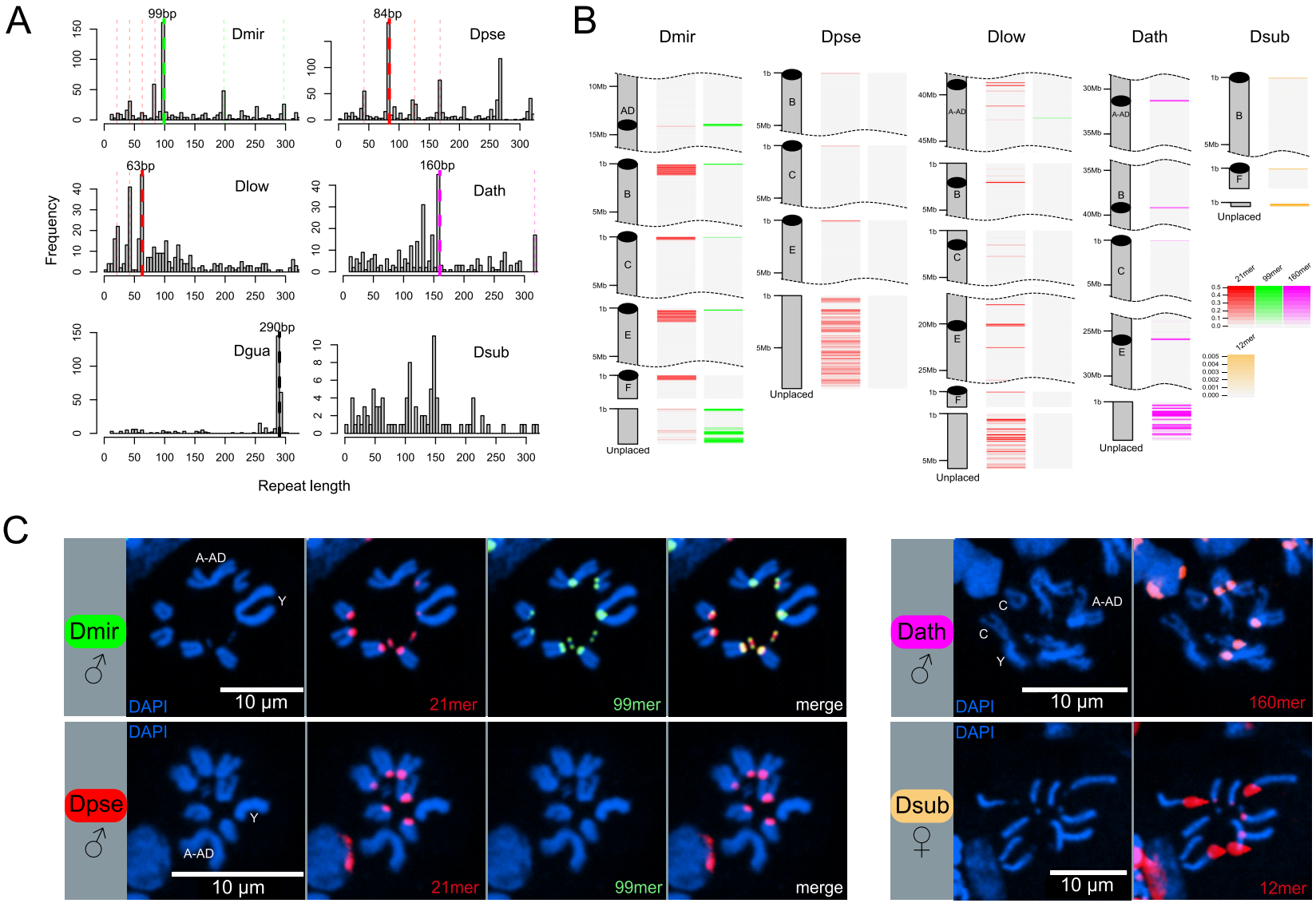
A

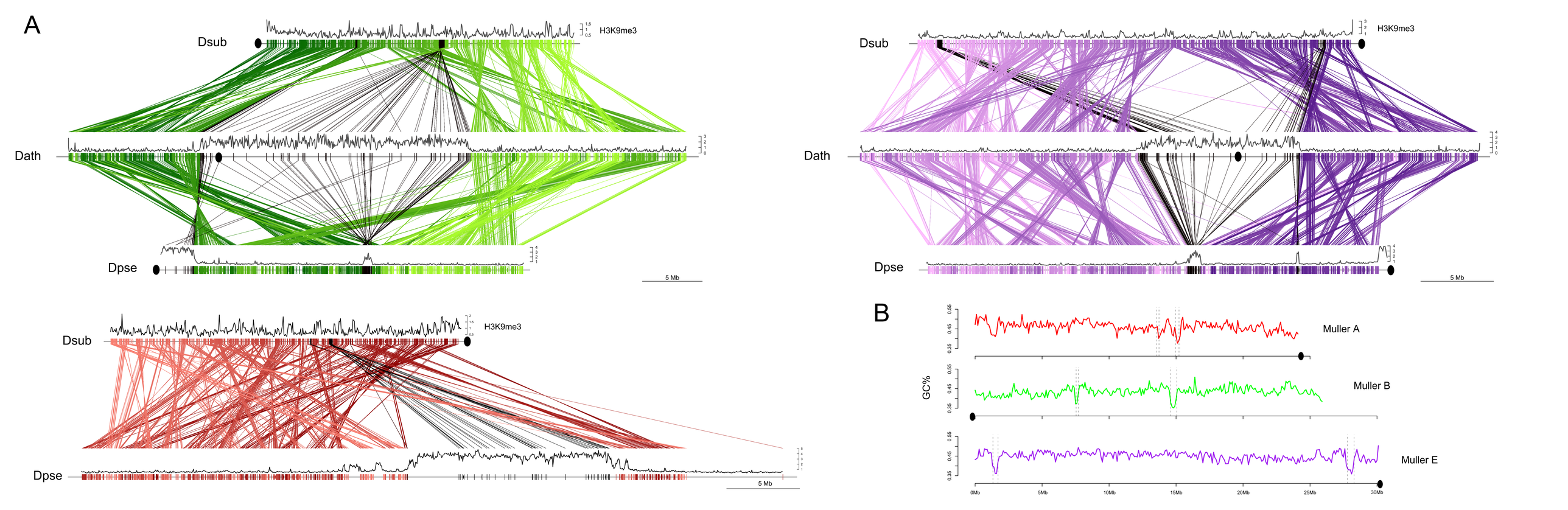

Dsub
Dath
A
B
C
D
E
F
A-AD
B
C
E
F
0Mb
10Mb
20Mb
30Mb
40Mb
50Mb
60Mb
70Mb
Dmel
X
2L
2R
3L
3R
4
Dlow
Dmir
Dpse

B


Muller A
Muller D
Muller C
Dsub
Dath
Dlow
Dmir
Dpse


Muller B
Muller E
Dsub
Dath
Dlow
Dmir
Dpse
10Mb


A
99bp
Dmir
84bp
Dpse
63bp
Dlow
160bp
Dath
290bp
Dgua
Dsub
Frequency
Repeat length
B
Dmir
Dpse
Dlow
Dath
Dsub
AD
A-AD
B
C
E
F
Unplaced
21mer
99mer
160mer
12mer
C
A-AD
Dath
Dmir
Dpse
Dsub
DAPI
10 μm
21mer
99mer
merge
160mer
12mer
Y


A
Dsub
H3K9me3
Dath
Dpse
5 Mb
Dsub
H3K9me3
Dath
Dpse
5 Mb
Dsub
H3K9me3
Dpse
5 Mb
B
GC%
Muller A
Muller B
Muller E
0Mb
5Mb
10Mb
15Mb
20Mb
25Mb
30Mb

A

telocentric to metacentric

metacentric to telocentric

pericentric inversions

pericentromere repositioning

inversion

A B C D

A B D C

B

inversion model

repositioning model

centromere

pericentromere

pericentric gene

C

telocentric to metacentric

metacentric to telocentric

Muller A (A-AD)

Muller B

Muller E

Dpse

Dath

Dlow

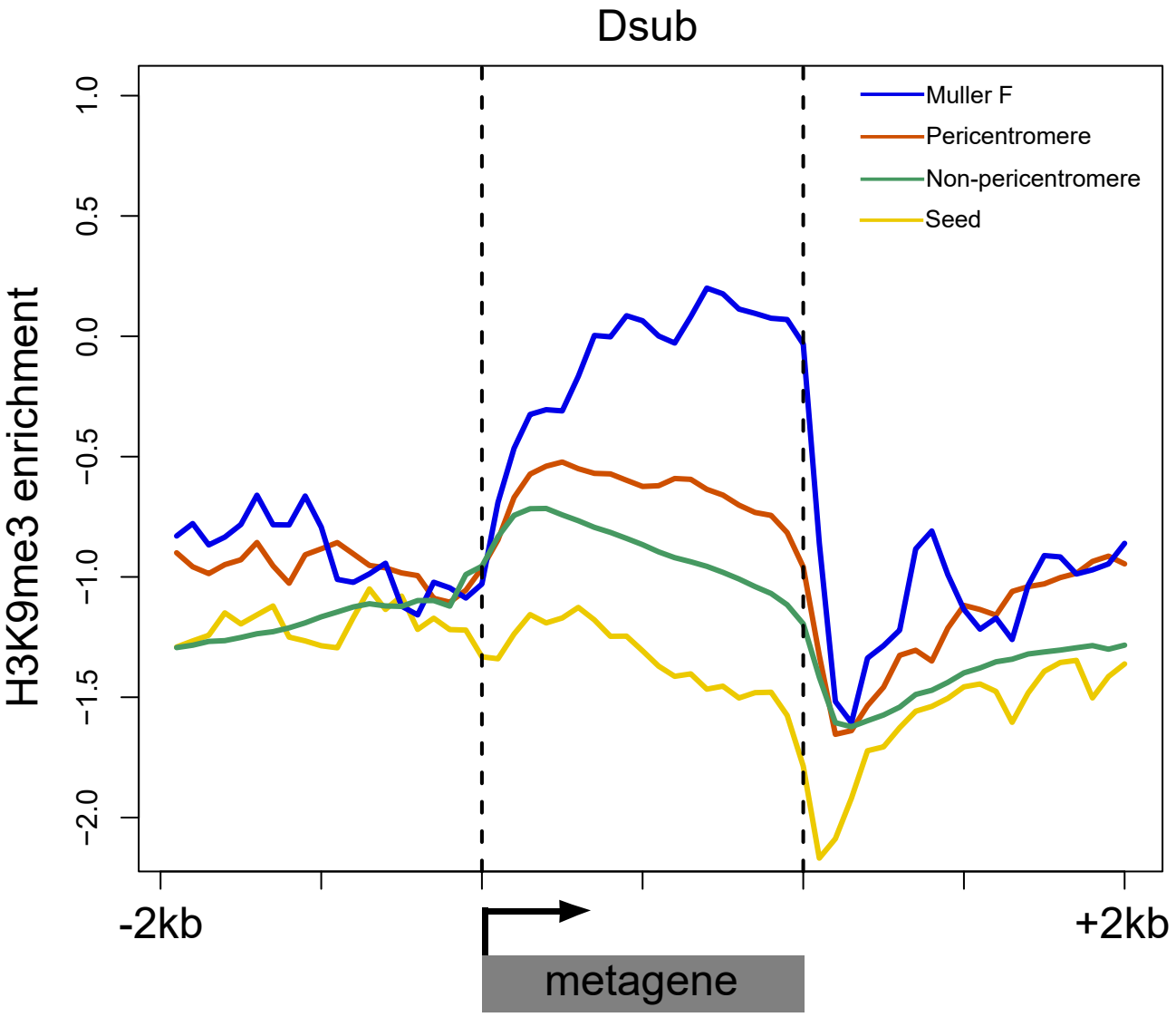
Dsub

Dlow

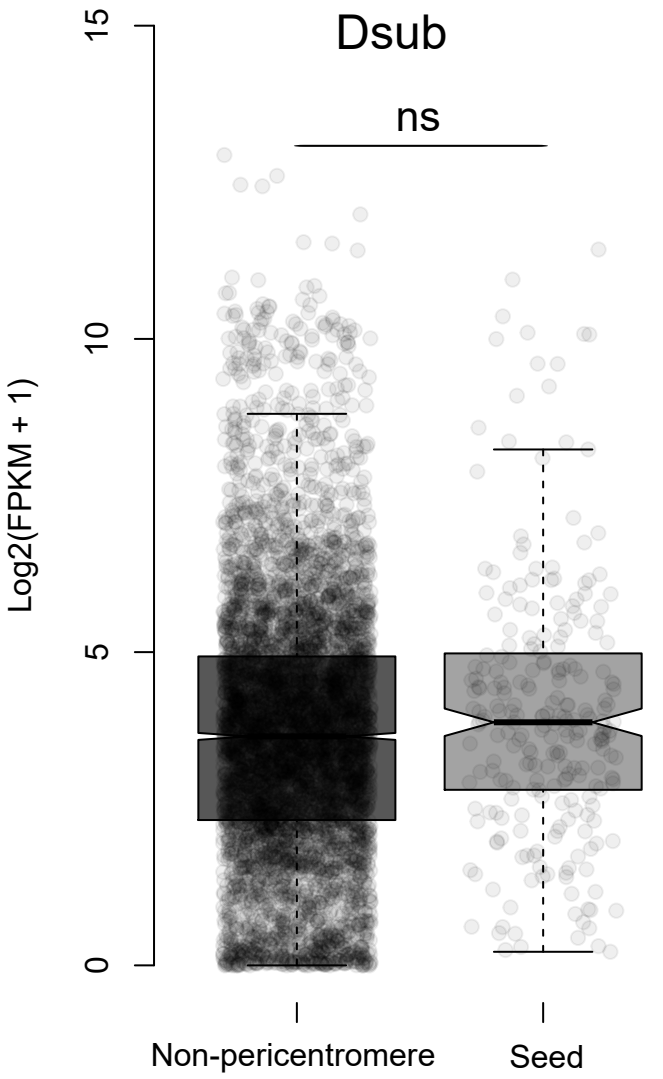
Dmir

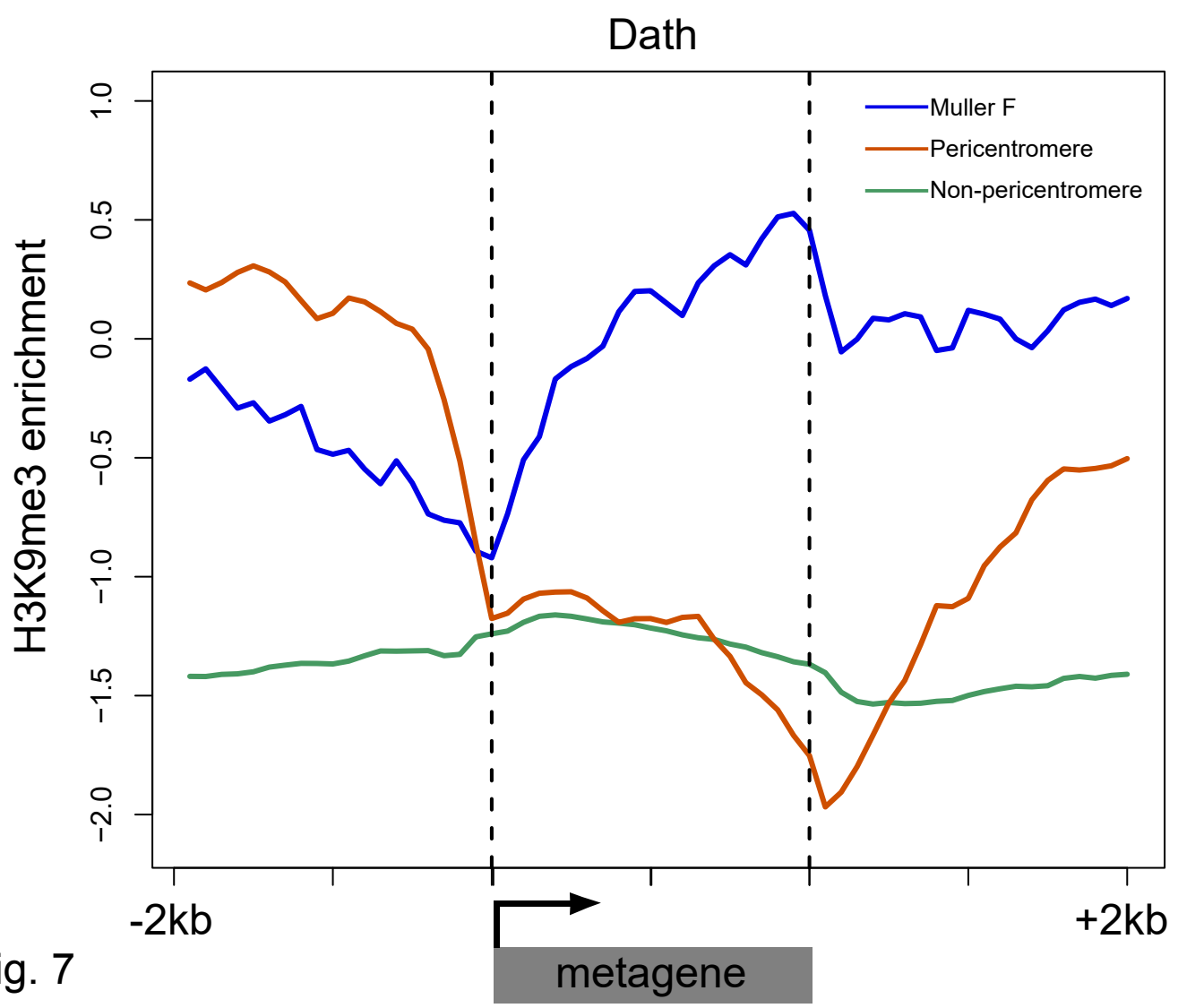
0Mb 10Mb 20Mb 30Mb 40Mb 50Mb

A


Dsub
H3K9me3 enrichment
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Muller F
Pericentromere
Non-pericentromere
Seed
-2kb
+2kb
metagene

B


Dsub
ns
Log2(FPKM + 1)
15
10
5
0
Non-pericentromere
Seed


Dath
H3K9me3 enrichment
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Muller F
Pericentromere
Non-pericentromere
-2kb
+2kb
metagene

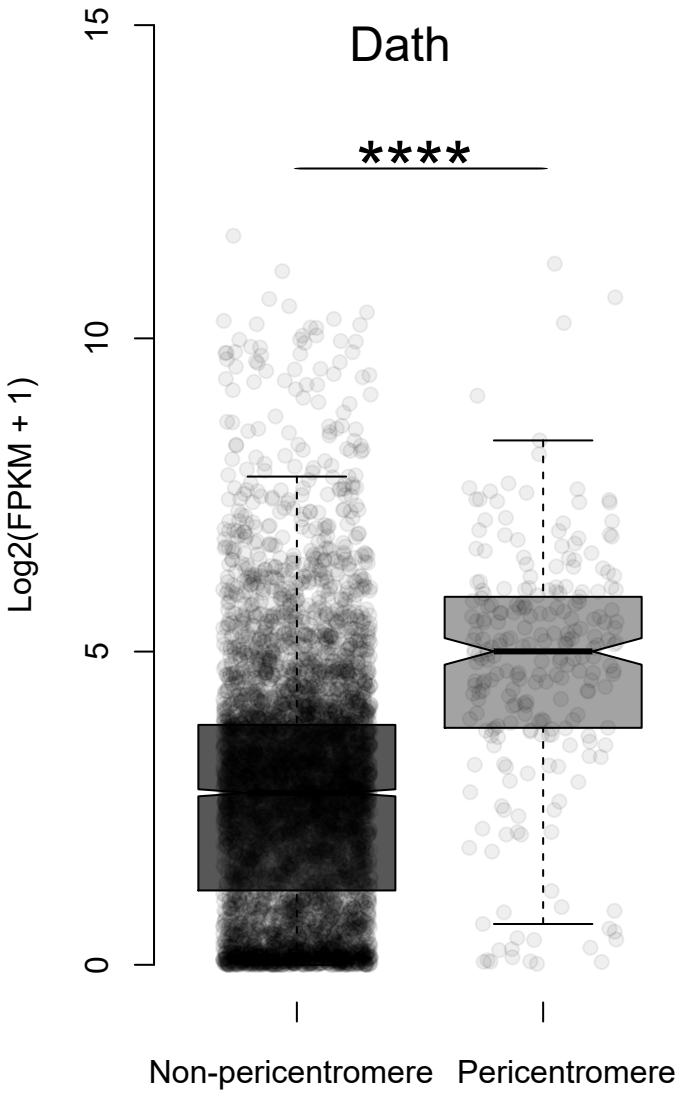
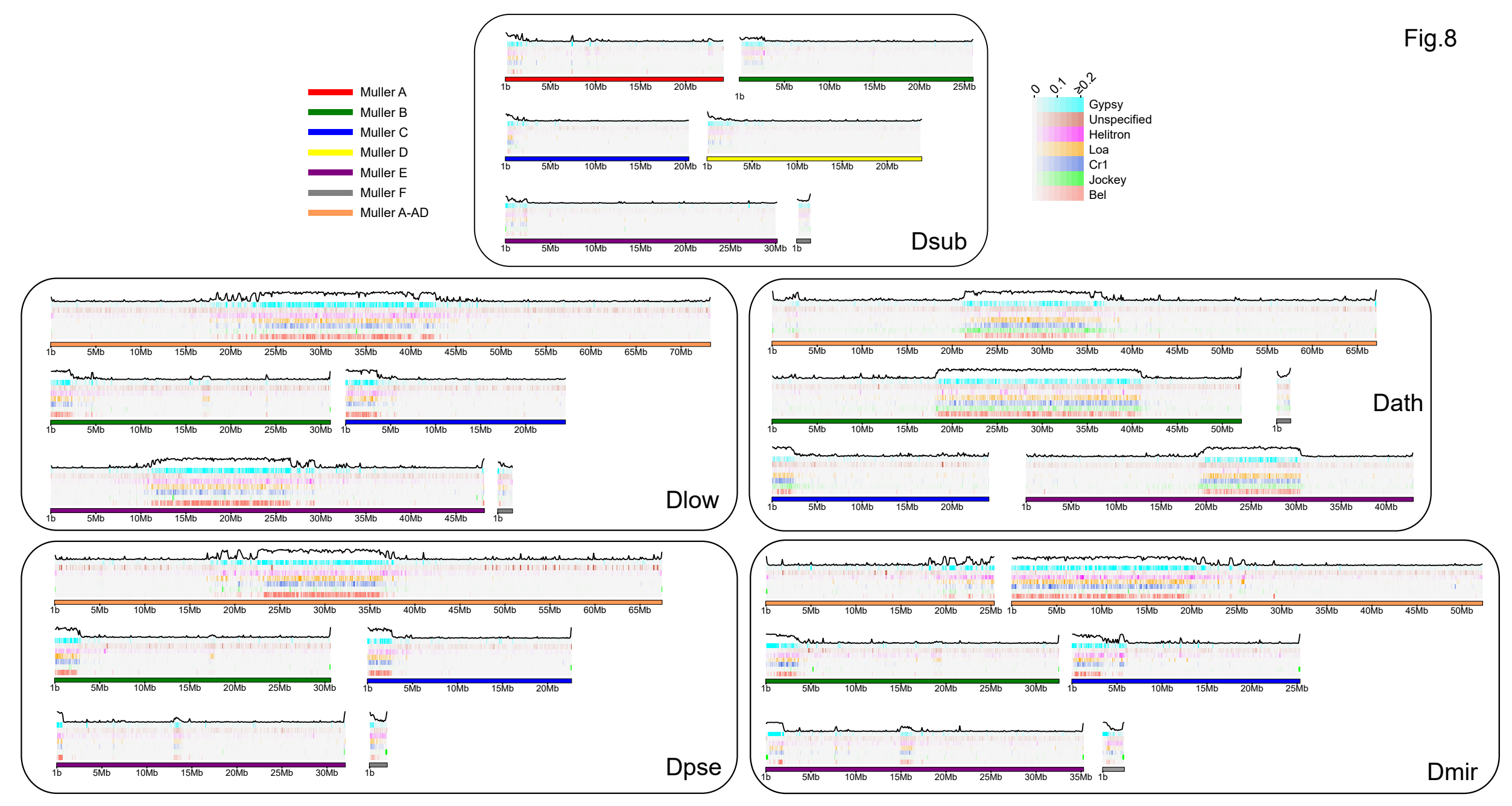
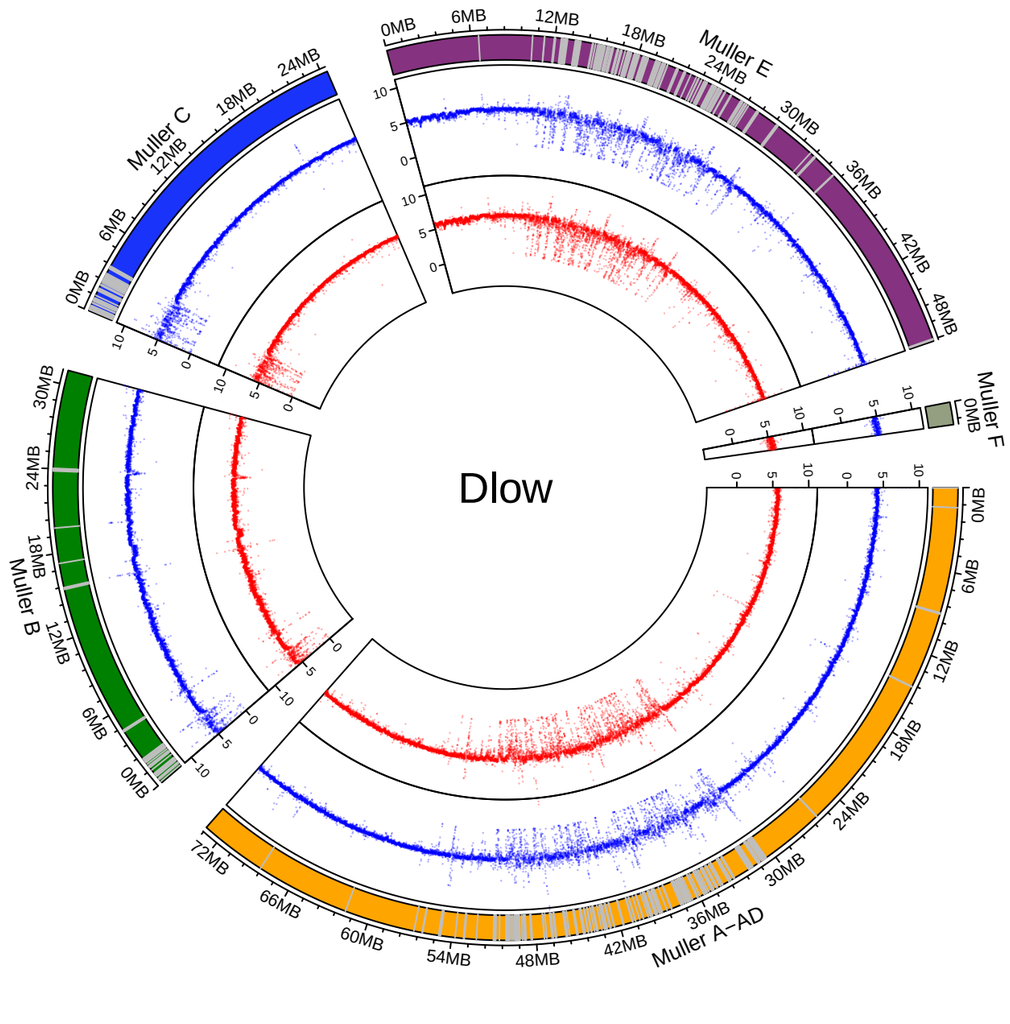
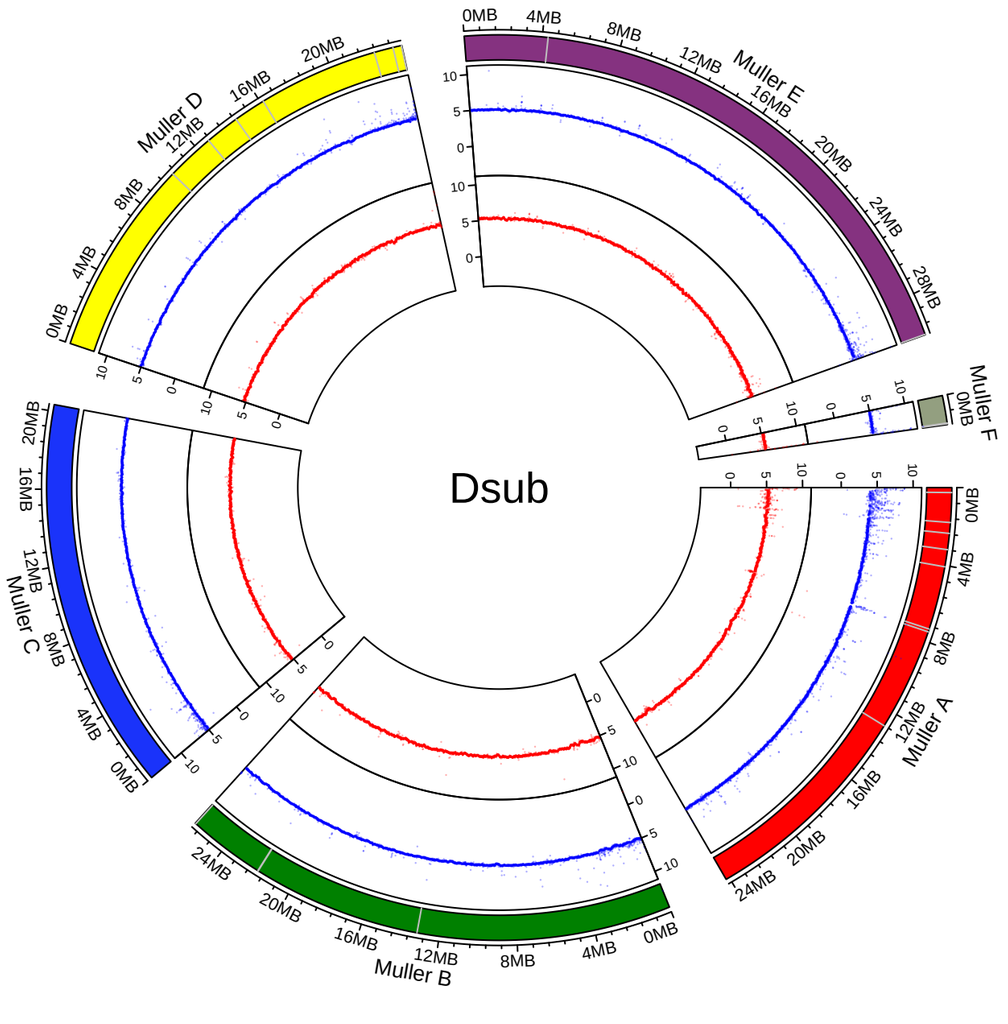

Dath
****
Log2(FPKM + 1)
15
10
5
0
Non-pericentromere
Pericentromere

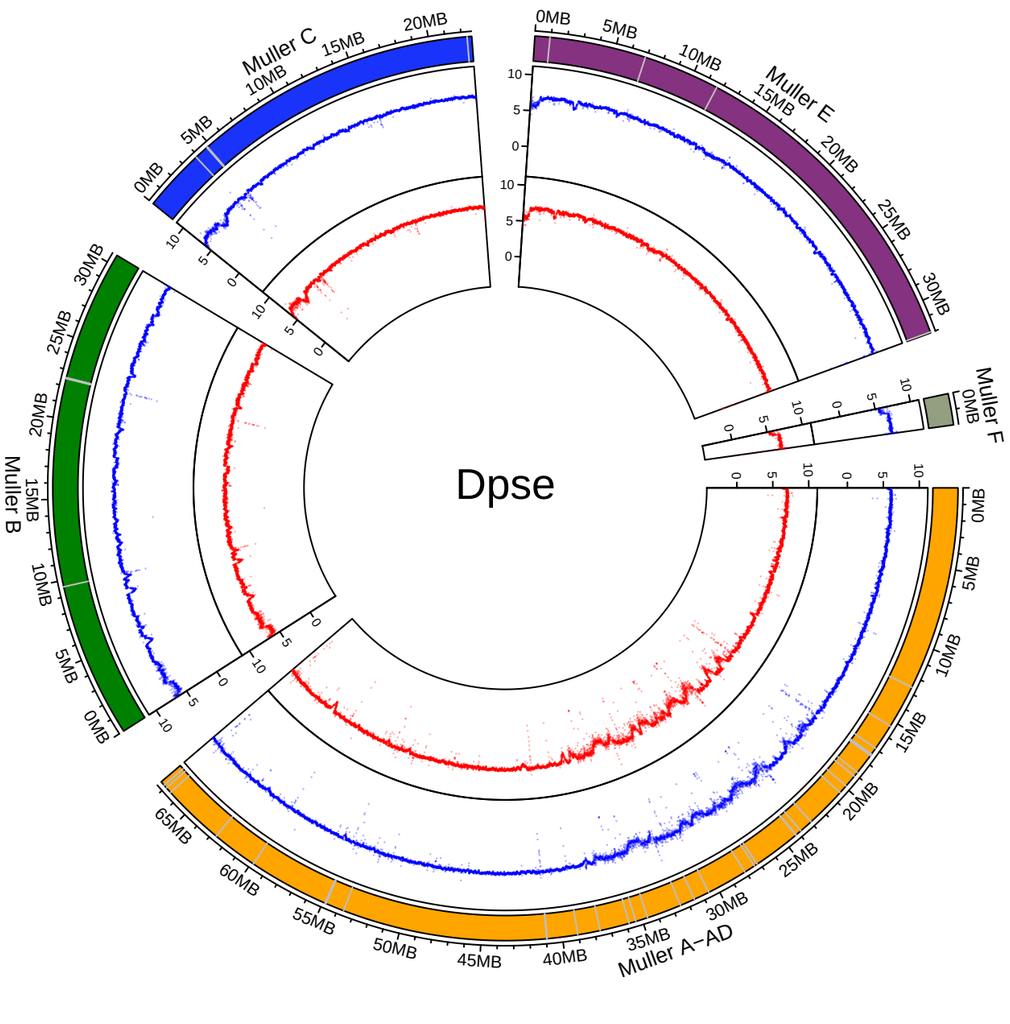
Fig. 7

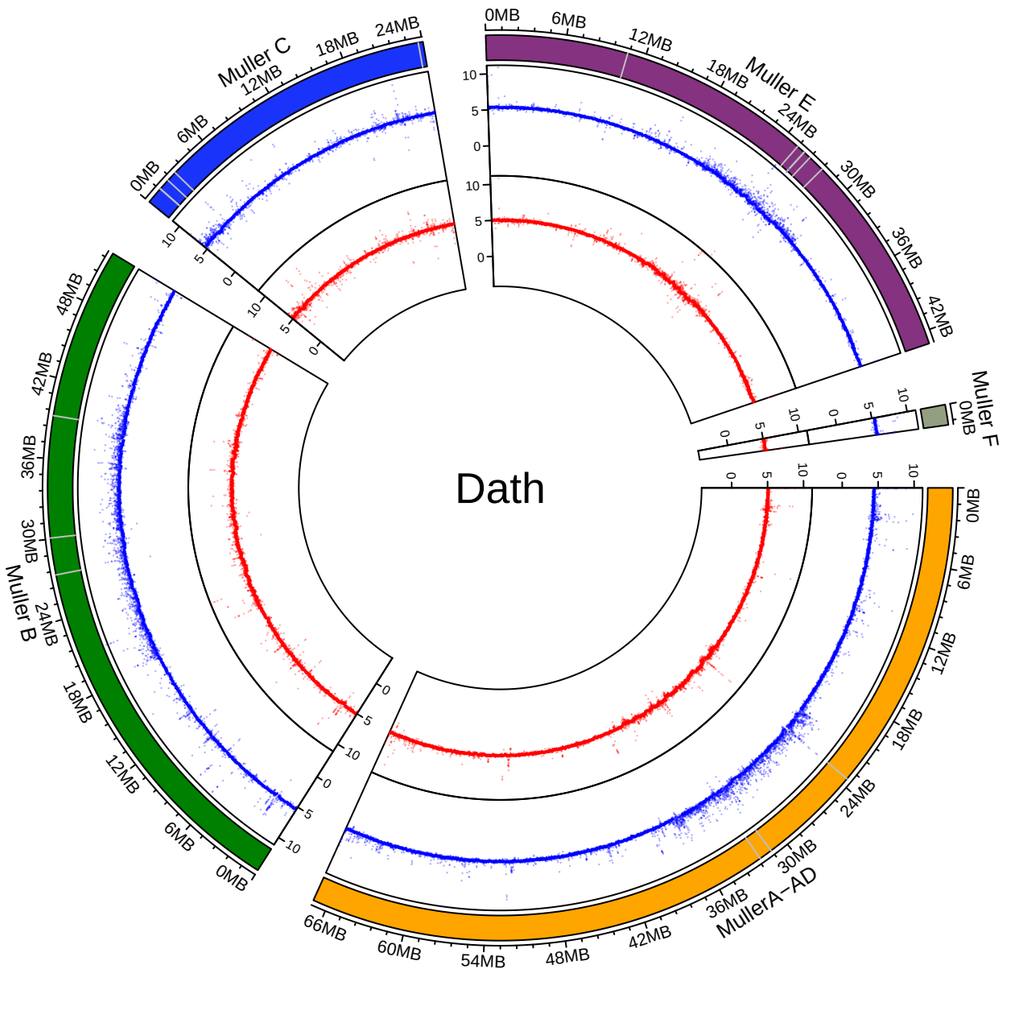
Fig.8

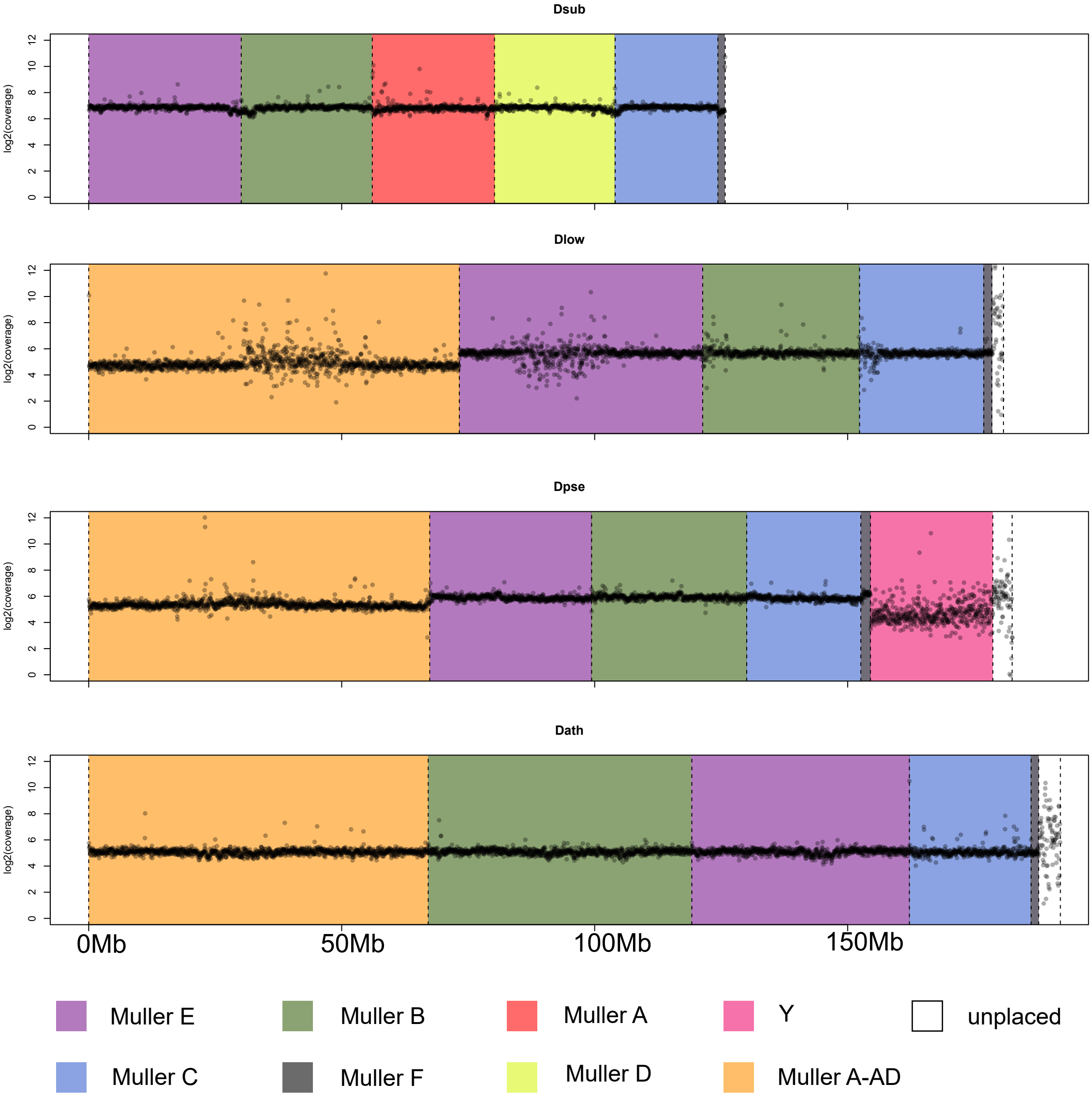

Muller A
Muller B
Muller C
Muller D
Muller E
Muller F
Muller A-AD
0
0.1
≥0.2
Gypsy
Unspecified
Helitron
Loa
Cr1
Jockey
Bel
Dsub
Dlow
Dath
Dpse
Dmir


Dlow
Muller A–AD
Muller B
Muller C
Muller E
Muller F

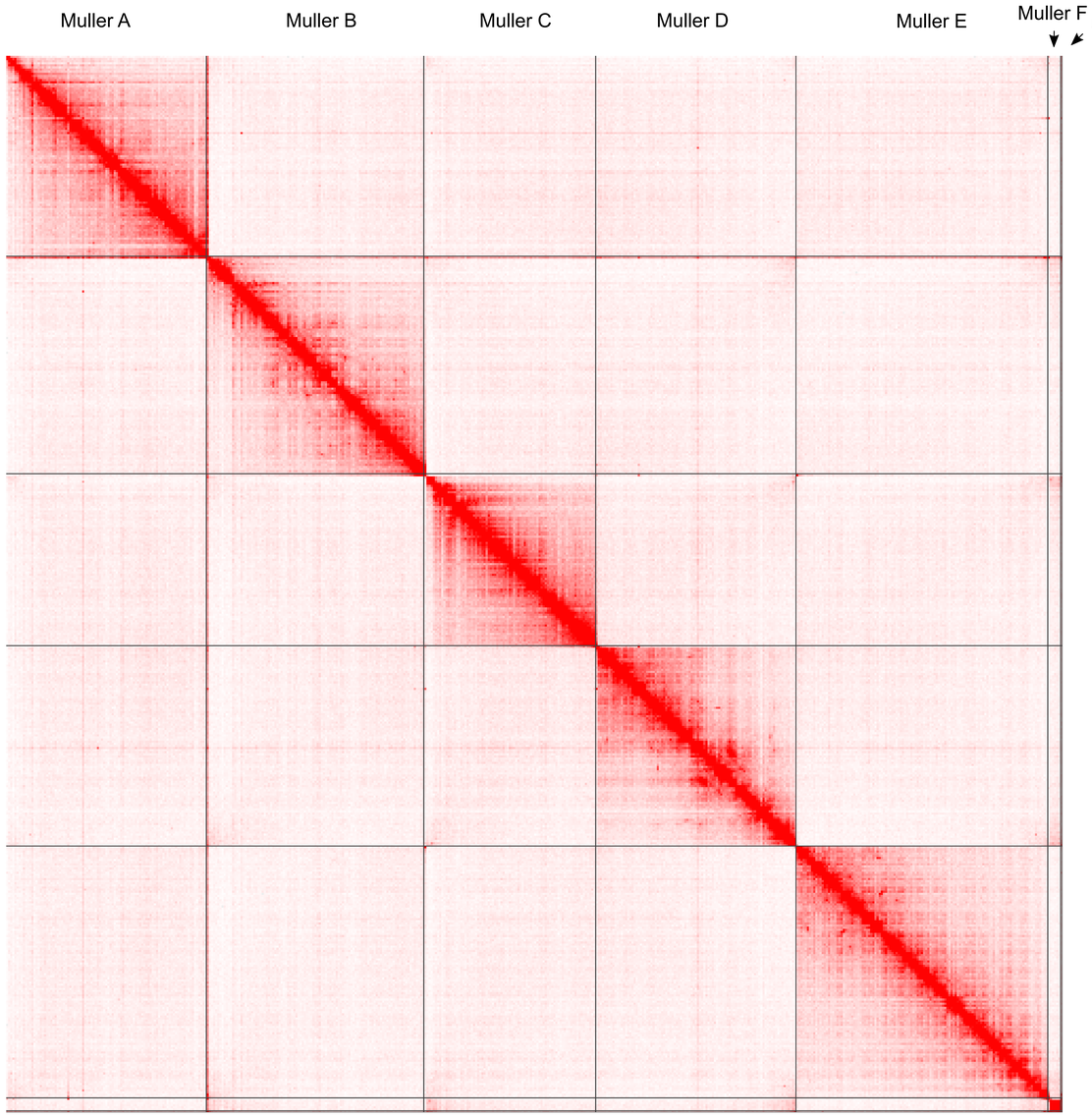

Dsub
Muller A
Muller B
Muller C
Muller D
Muller E
Muller F


Dpse
Muller A–AD
Muller B
Muller C
Muller E
Muller F

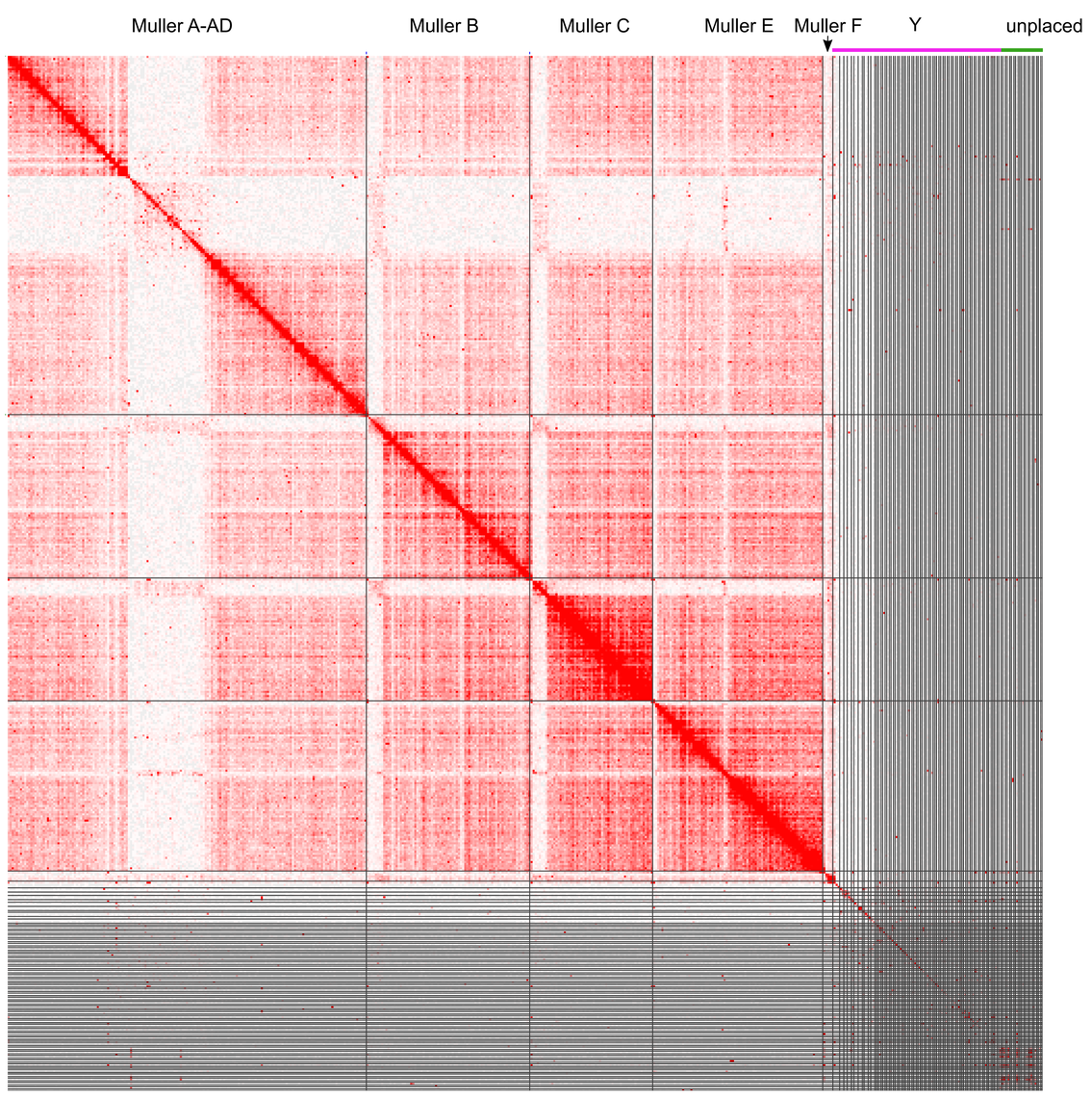

Dath
MullerA–AD
Muller B
Muller C
Muller E
Muller F


Dsub
log2(coverage)
0
2
4
6
8
10
12
Dlow
log2(coverage)
0
2
4
6
8
10
12
Dpse
log2(coverage)
0
2
4
6
8
10
12
Dath
log2(coverage)
0
2
4
6
8
10
12
0Mb
50Mb
100Mb
150Mb
Muller E
Muller B
Muller A
Y
unplaced
Muller C
Muller F
Muller D
Muller A-AD

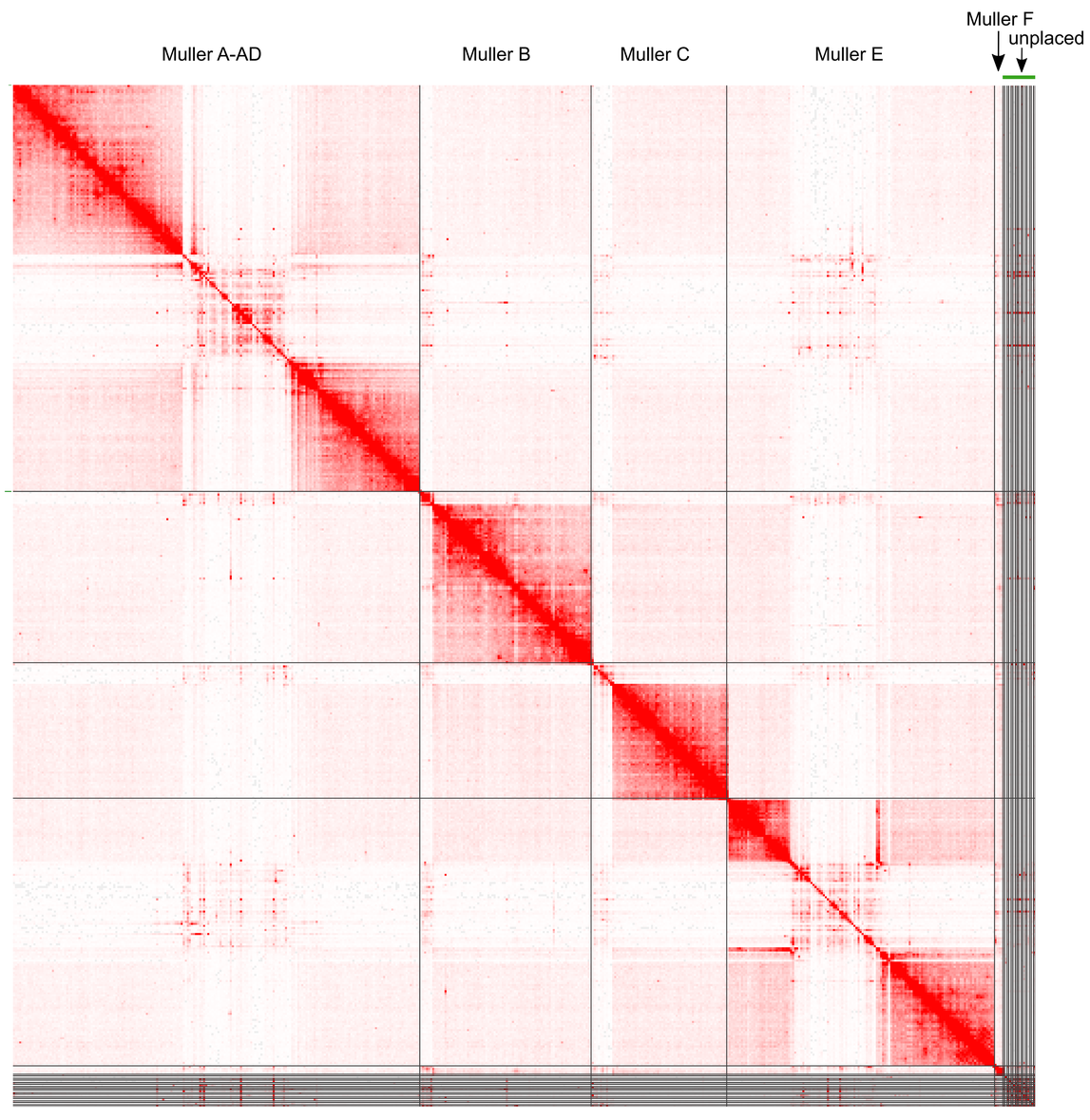
Dsub



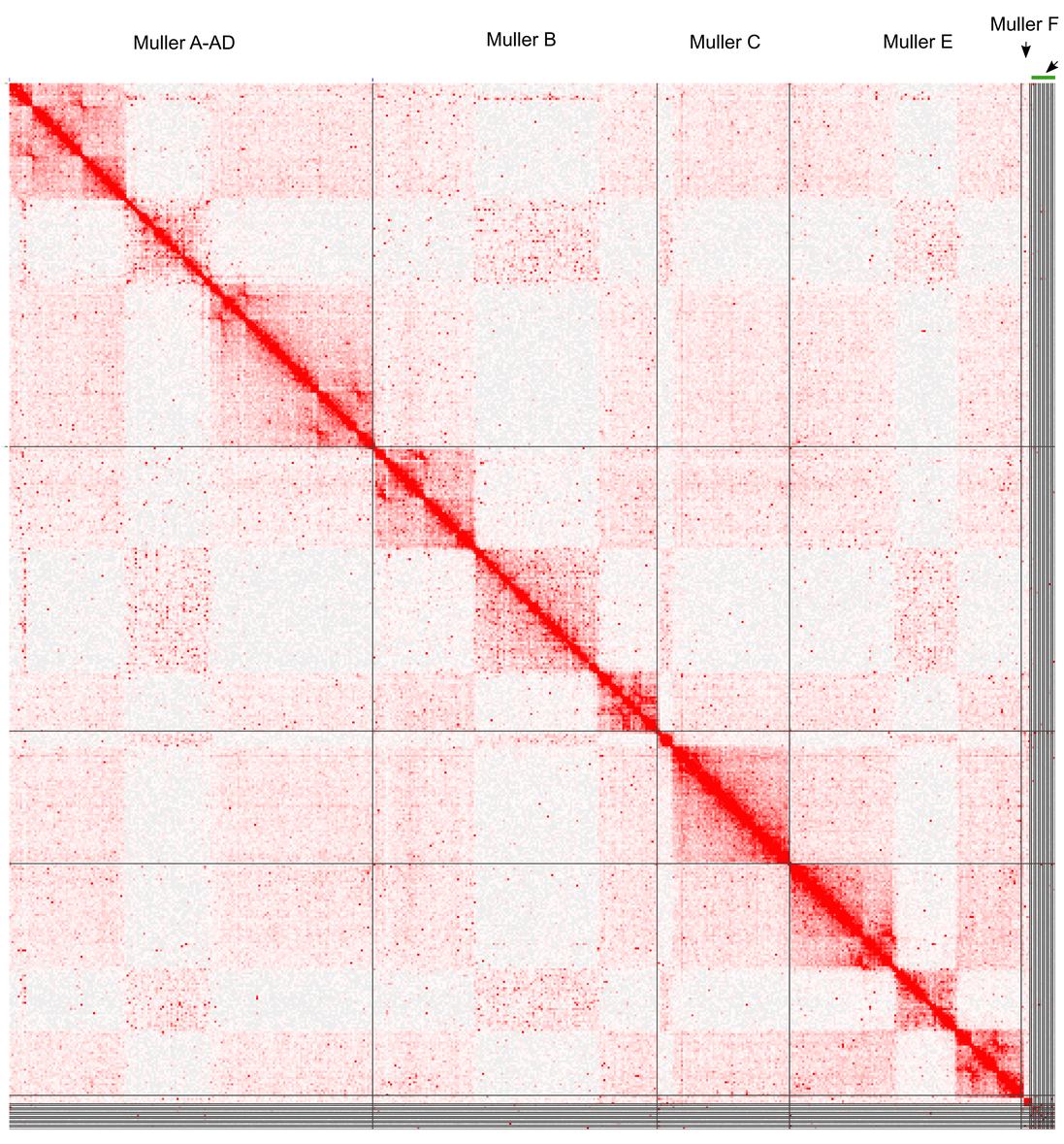
Dpse



Dlow

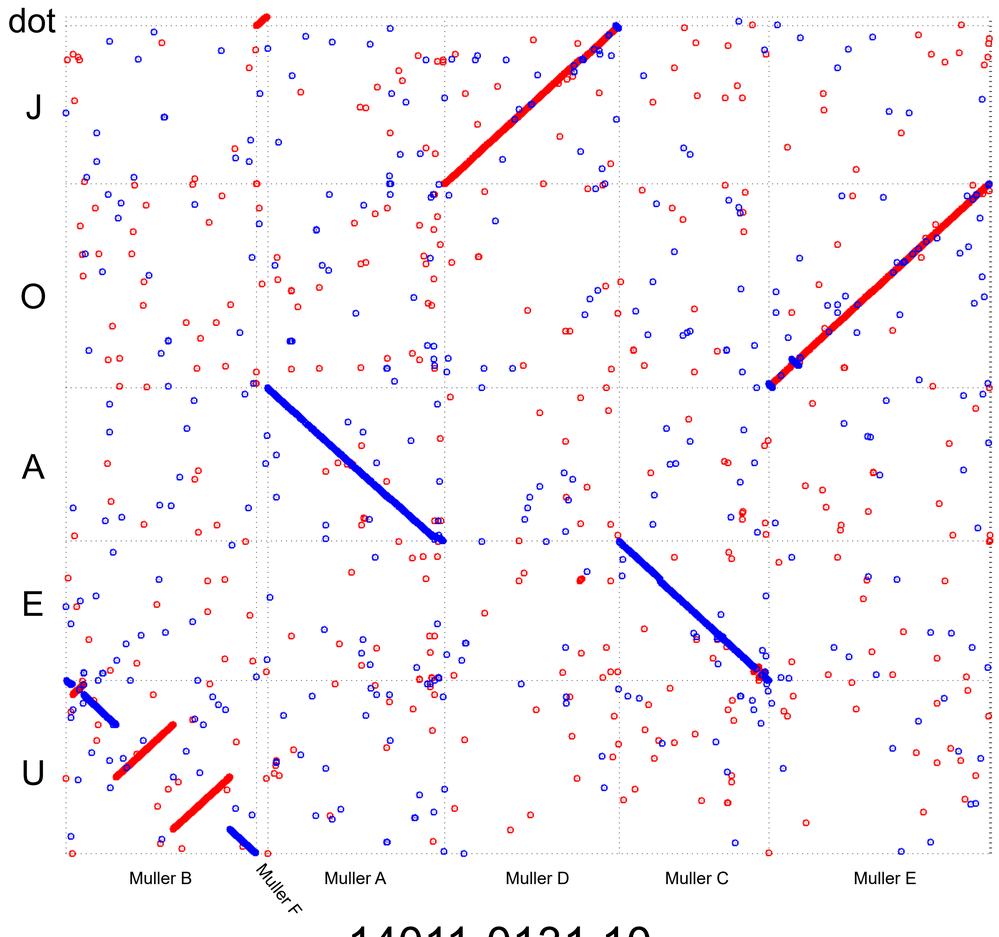


Dath EB hybrid



Dsub vs Dsub

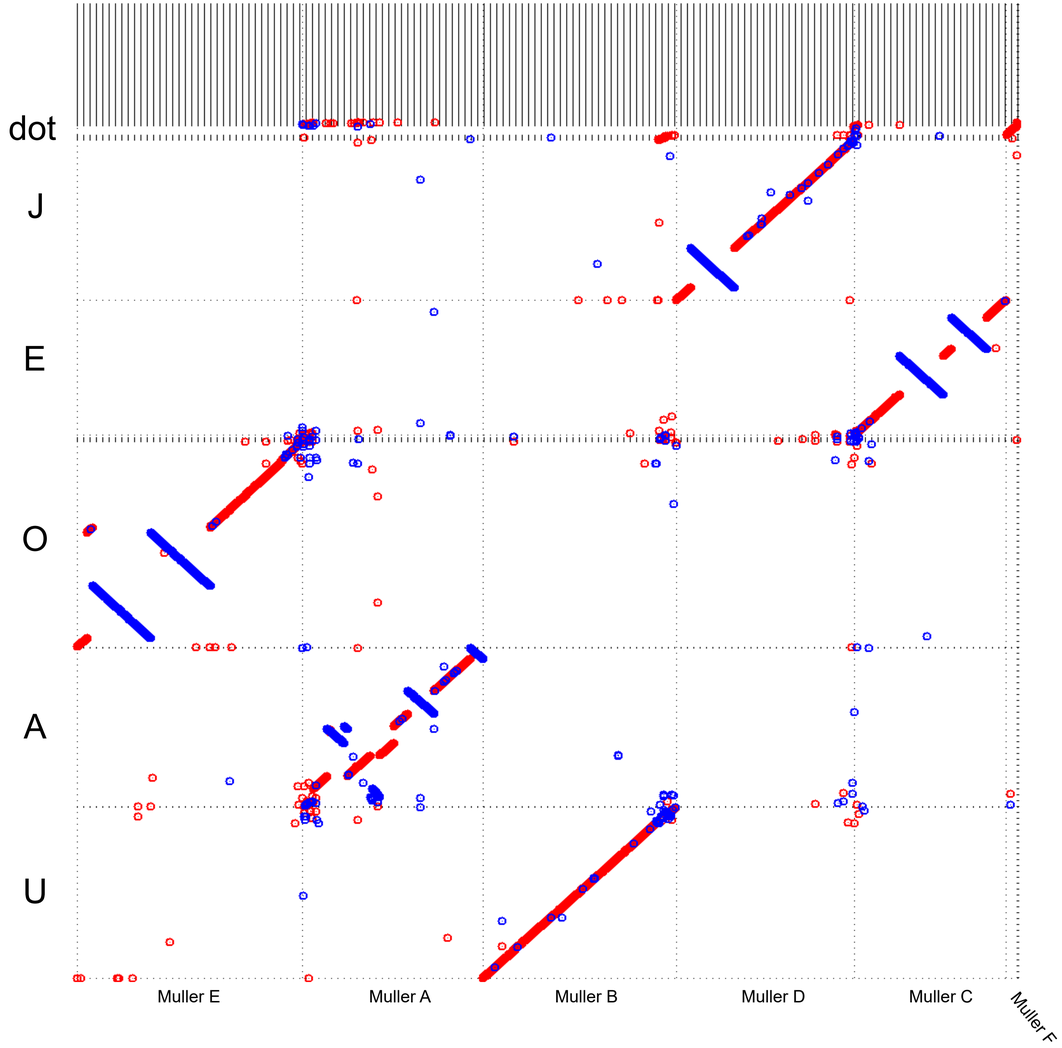
ch-cu



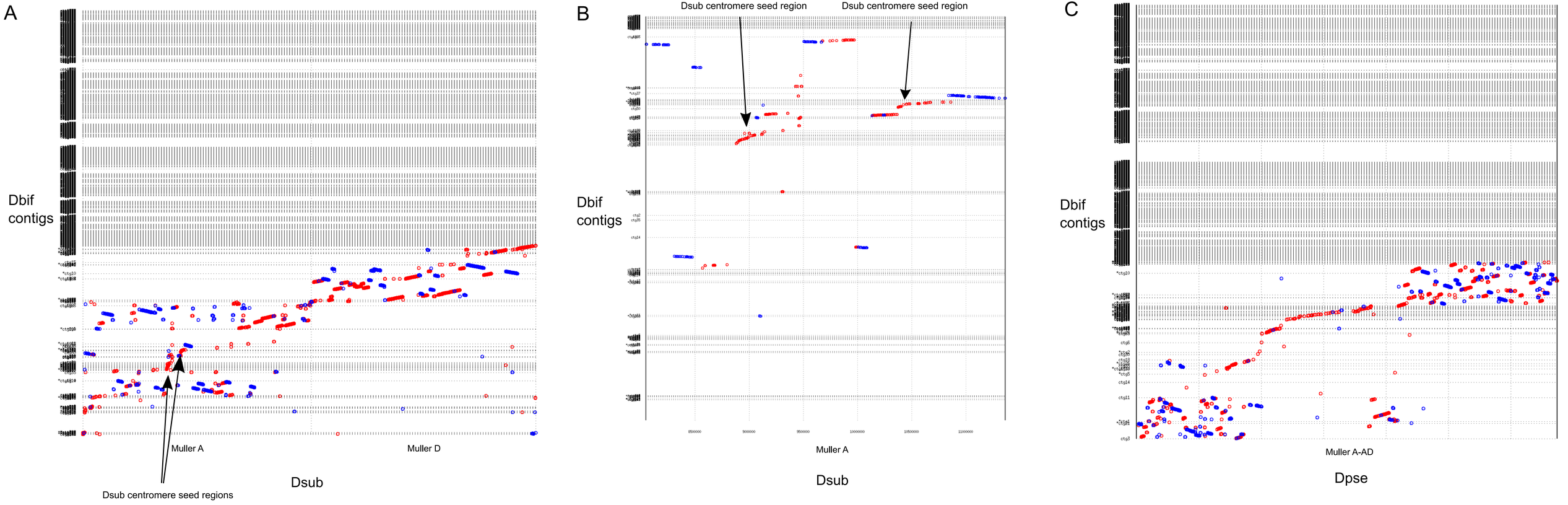
14011-0131.10

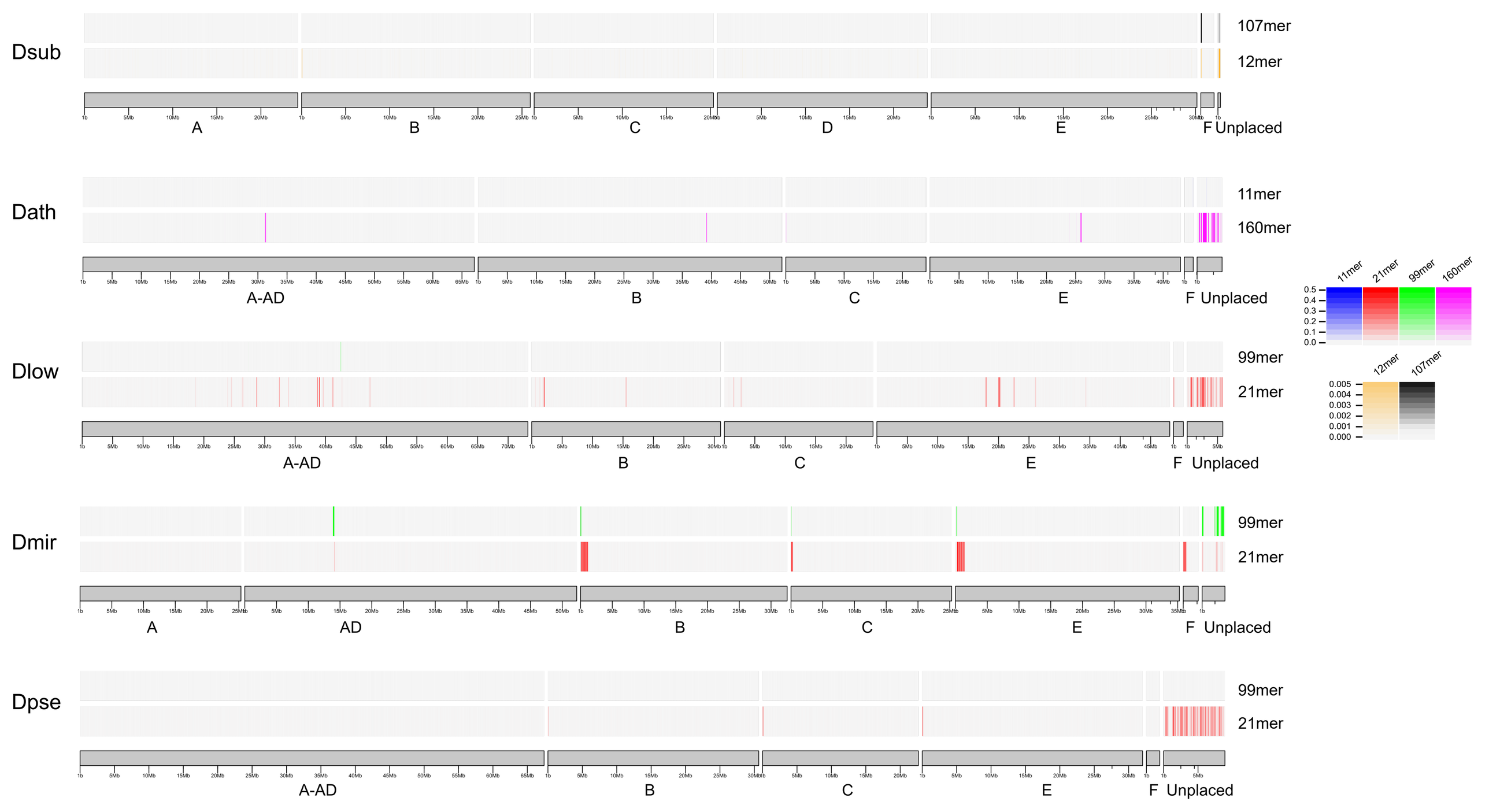
Dsub vs Dgua

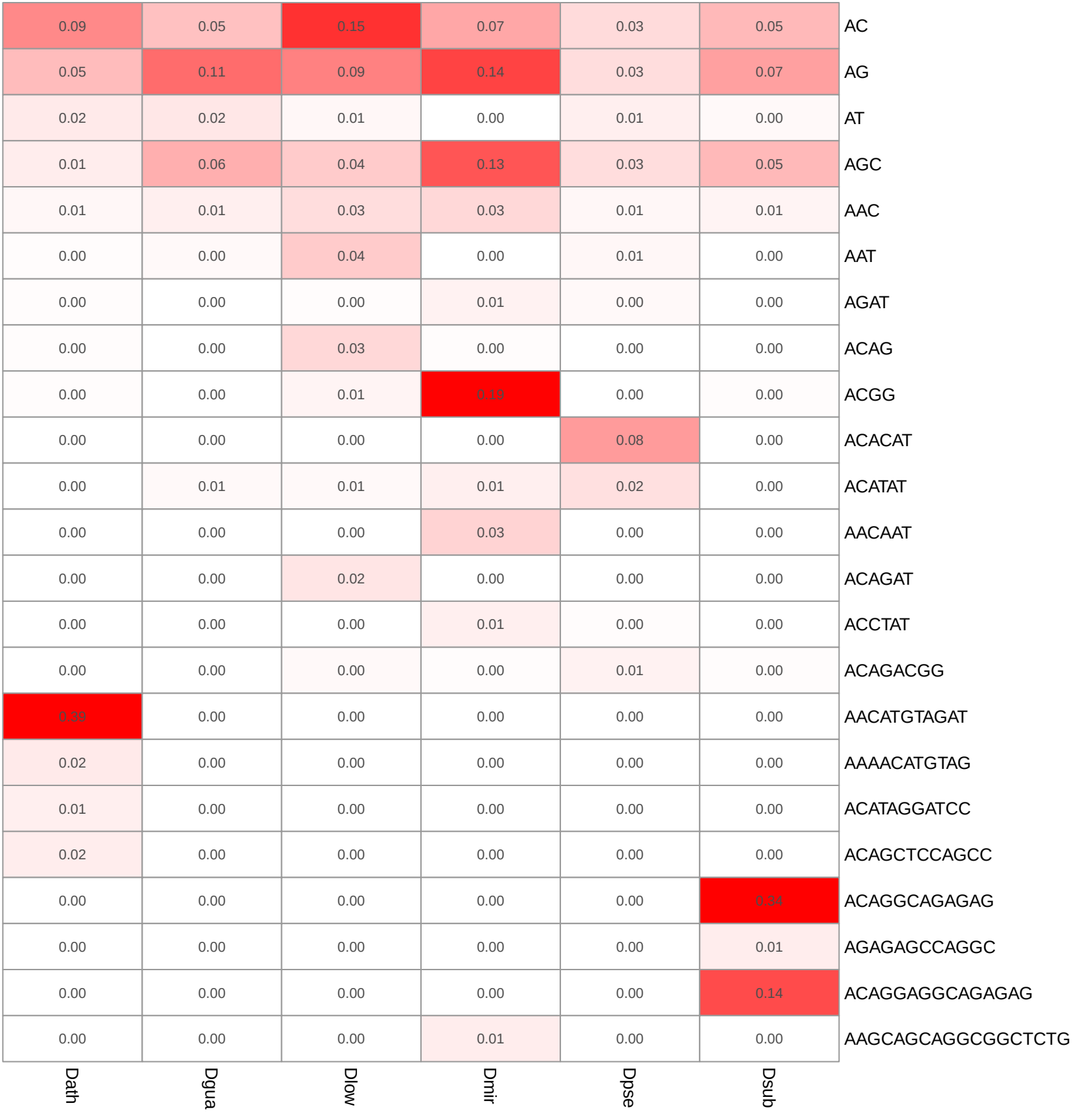
dgua6

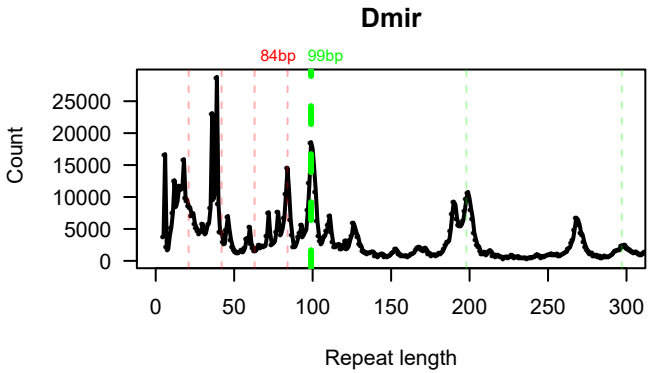


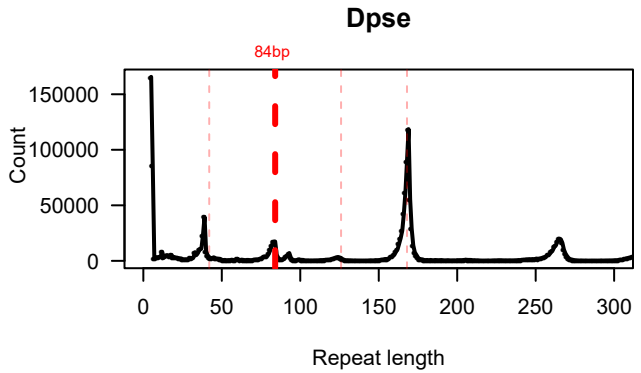
14011-0131.10

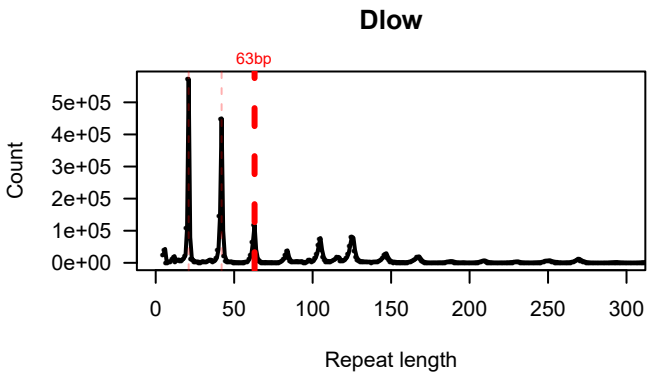

A
Dbif contigs
Muller A
Muller D
Dsub centromere seed regions
Dsub
B
Dsub centromere seed region
Dsub centromere seed region
Dbif contigs
Muller A
Dsub
C
Dbif contigs
Muller A-AD
Dpse

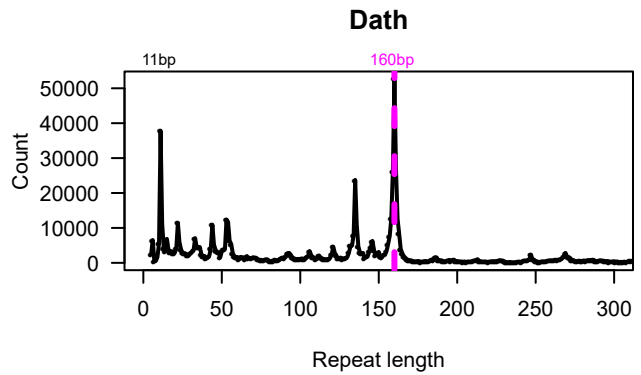

Dsub
107mer
12mer
A
B
C
D
E
F
Unplaced
Dath
11mer
160mer
A-AD
B
C
E
F
Unplaced
Dlow
99mer
21mer
A-AD
B
C
E
F
Unplaced
Dmir
99mer
21mer
A
AD
B
C
E
F
Unplaced
Dpse
99mer
21mer
A-AD
B
C
E
F
Unplaced
11mer
21mer
99mer
160mer
12mer
107mer

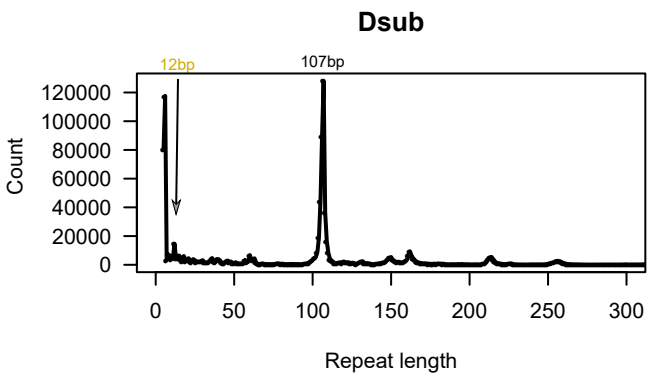

0.09 0.05 0.15 0.07 0.03 0.05 AC
0.05 0.11 0.09 0.14 0.03 0.07 AG
0.02 0.02 0.01 0.00 0.01 0.00 AT
0.01 0.06 0.04 0.13 0.03 0.05 AGC
0.01 0.01 0.03 0.03 0.01 0.01 AAC
0.00 0.00 0.04 0.00 0.01 0.00 AAT
0.00 0.00 0.00 0.01 0.00 0.00 AGAT
0.00 0.00 0.03 0.00 0.00 0.00 ACAG
0.00 0.00 0.01 0.19 0.00 0.00 ACGG
0.00 0.00 0.00 0.00 0.08 0.00 ACACAT
0.00 0.01 0.01 0.01 0.02 0.00 ACATAT
0.00 0.00 0.00 0.03 0.00 0.00 AACAAT
0.00 0.00 0.02 0.00 0.00 0.00 ACAGAT
0.00 0.00 0.00 0.01 0.00 0.00 ACCTAT
0.00 0.00 0.00 0.00 0.01 0.00 ACAGACGG
0.39 0.00 0.00 0.00 0.00 0.00 AACATGTAGAT
0.02 0.00 0.00 0.00 0.00 0.00 AAAACATGTAG
0.01 0.00 0.00 0.00 0.00 0.00 ACATAGGATCC
0.02 0.00 0.00 0.00 0.00 0.00 ACAGCTCCAGCC
0.00 0.00 0.00 0.00 0.00 0.34 ACAGGCAGAGAG
0.00 0.00 0.00 0.00 0.00 0.01 AGAGAGCCAGGC
0.00 0.00 0.00 0.00 0.00 0.14 ACAGGAGGCAGAGAG
0.00 0.00 0.00 0.01 0.00 0.00 AAGCAGCAGGCGGCTCTG
Dath
Dgua
Dlow
Dmir
Dpse
Dsub


Dmir
84bp
99bp
Count
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300
Repeat length

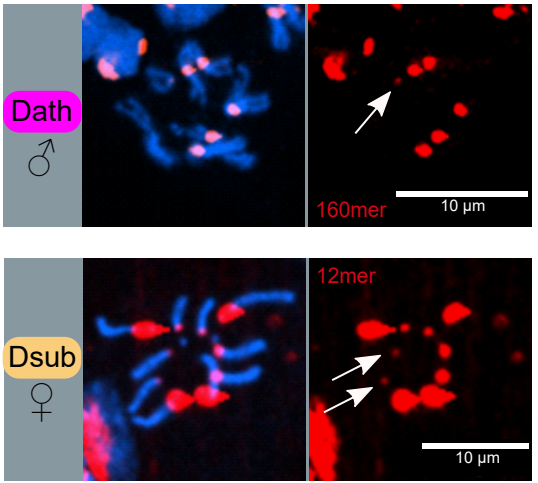

Dpse
84bp
Count
150000
100000
50000
0
0
50
100
150
200
250
300
Repeat length


Dlow
63bp
Count
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
50
100
150
200
250
300
Repeat length

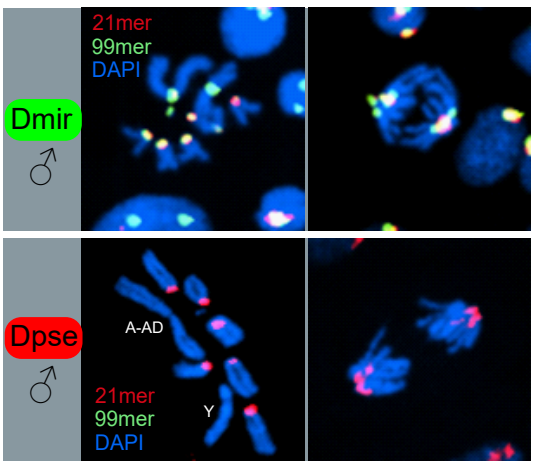

Dath
11bp
160bp
Count
50000
40000
30000
20000
10000
0
0
50
100
150
200
250
300
Repeat length

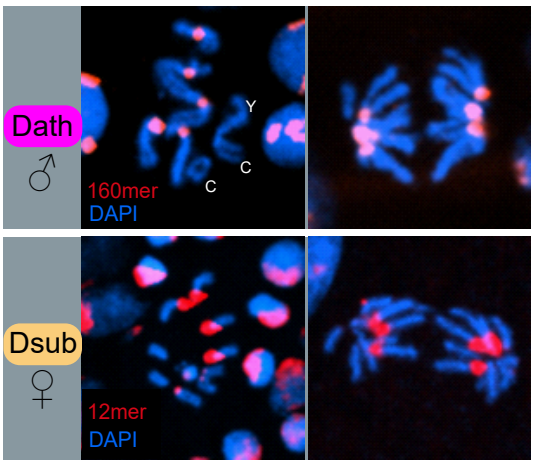

Dsub
12bp
107bp
Count
120000
100000
80000
60000
40000
20000
0
0
50
100
150
200
250
300
Repeat length

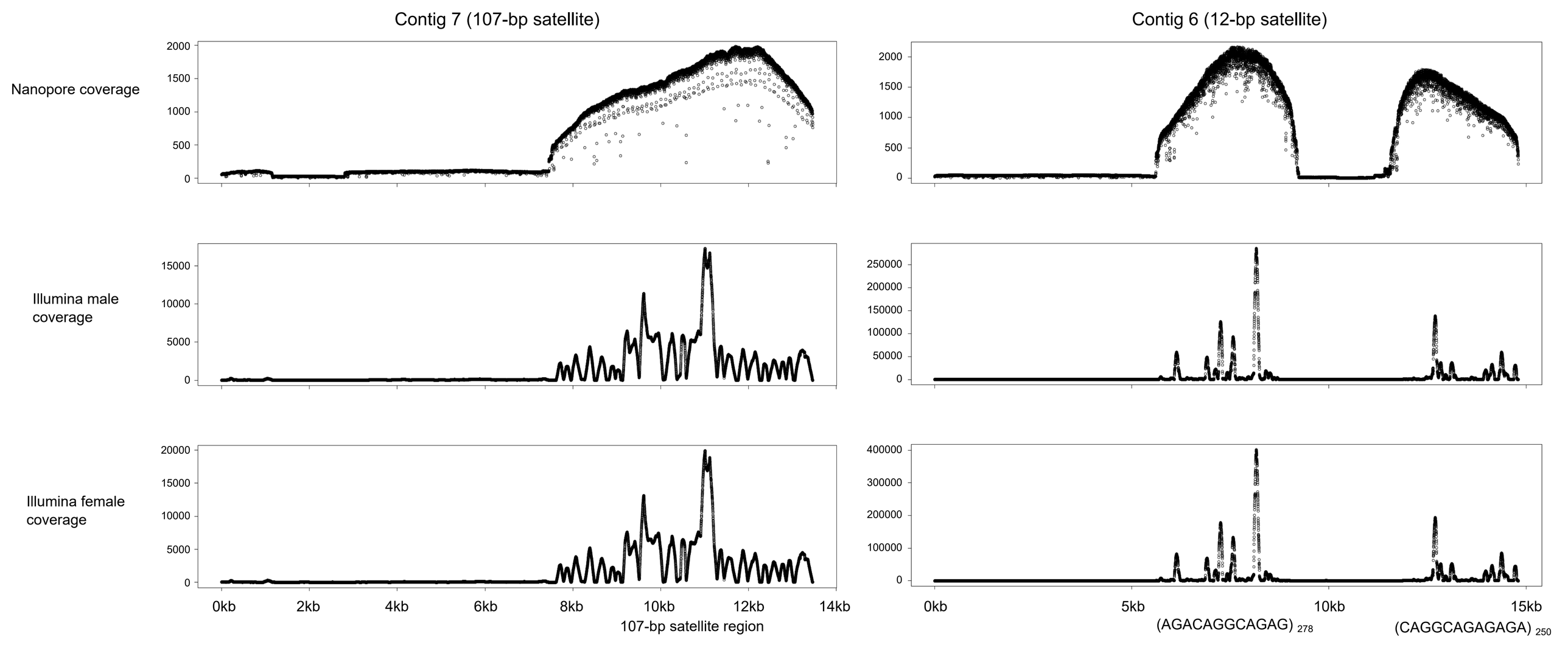
A

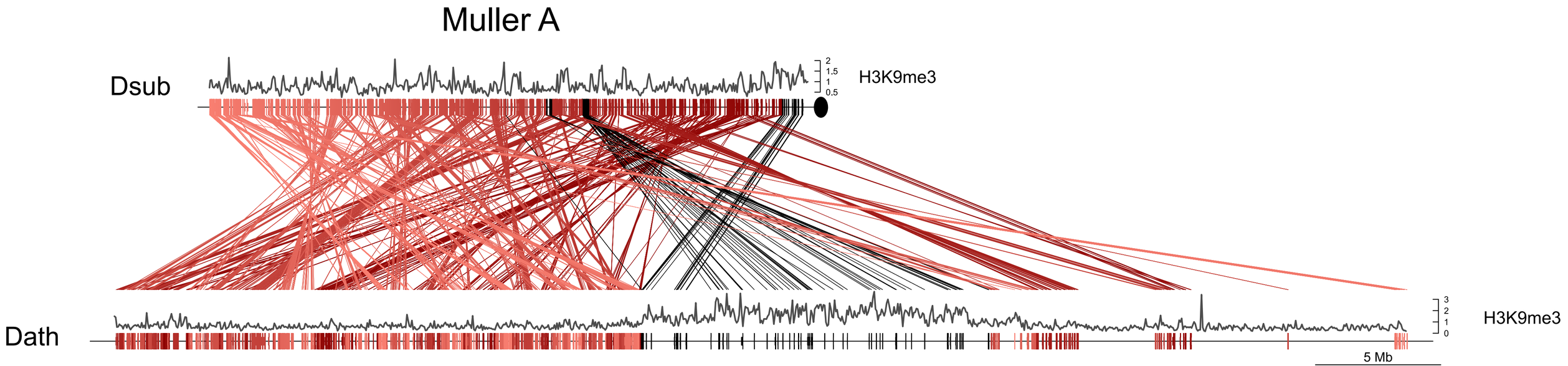

Dath
♂
160mer
10 μm
Dsub
♀
12mer
10 μm

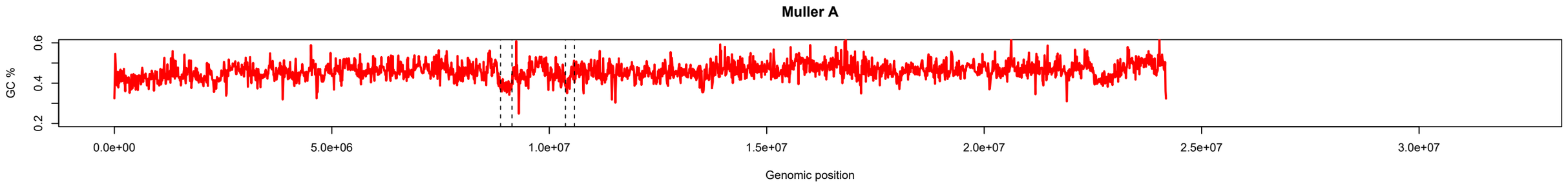
B

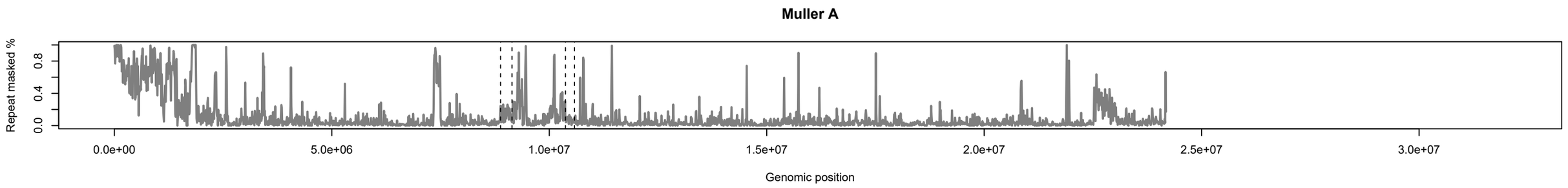

Dmir
♂
21mer
99mer
DAPI
Dpse
♂
A-AD
21mer
99mer
DAPI
Y

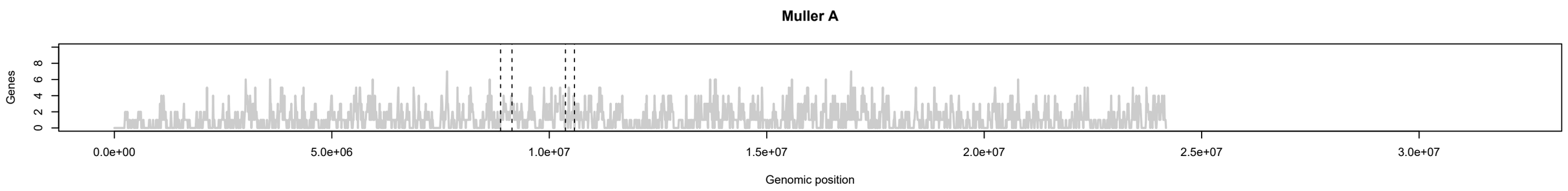

Dath
♂
Y
C
C
160mer
DAPI
Dsub
♀
12mer
DAPI

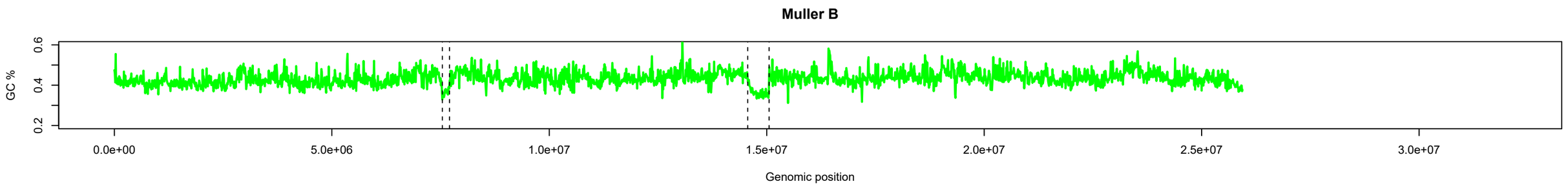

Contig 7 (107-bp satellite)
Contig 6 (12-bp satellite)
Nanopore coverage
Illumina male coverage
Illumina female coverage
2000
1500
1000
500
0
15000
10000
5000
20000
250000
200000
150000
100000
50000
400000
300000
0kb
2kb
4kb
6kb
8kb
10kb
12kb
14kb
107-bp satellite region
5kb
15kb
(AGACAGGCAGAG) 278
(CAGGCAGAGAGA) 250

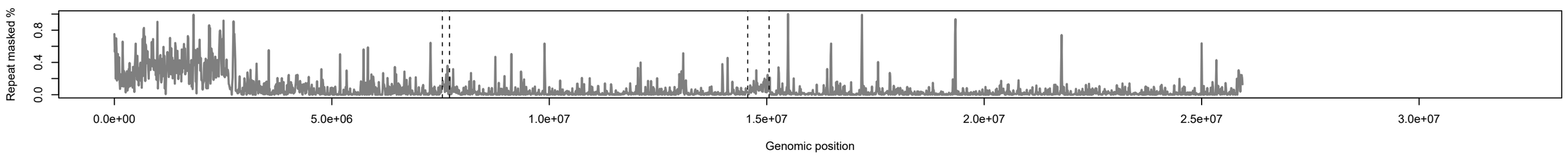

Muller A
Dsub
2
1.5
1
0.5
H3K9me3
Dath
3
2
1
0
H3K9me3
5 Mb

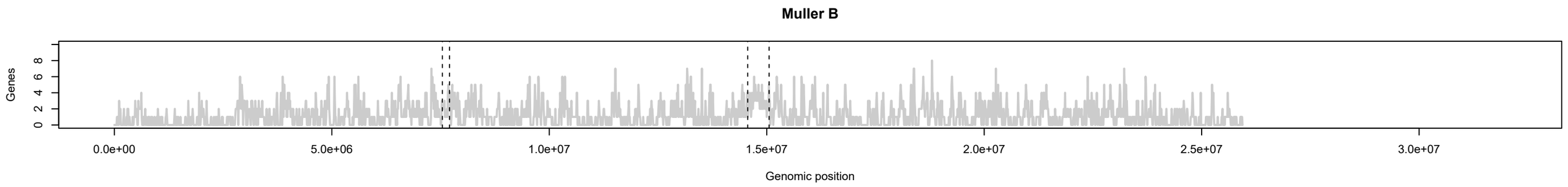

Muller A
GC %
0.2
0.4
0.6
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

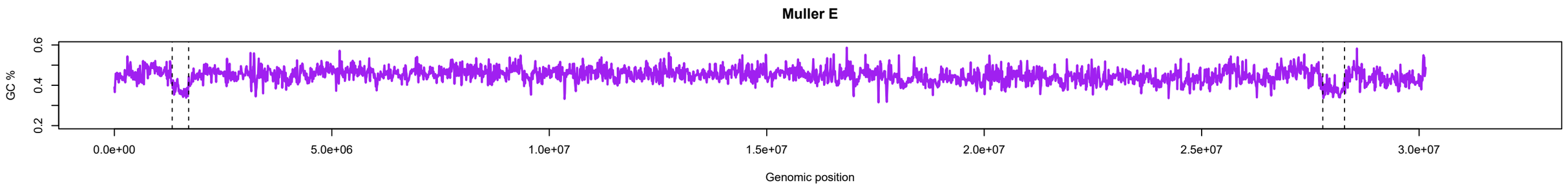

Muller A
Repeat masked %
0.0
0.4
0.8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

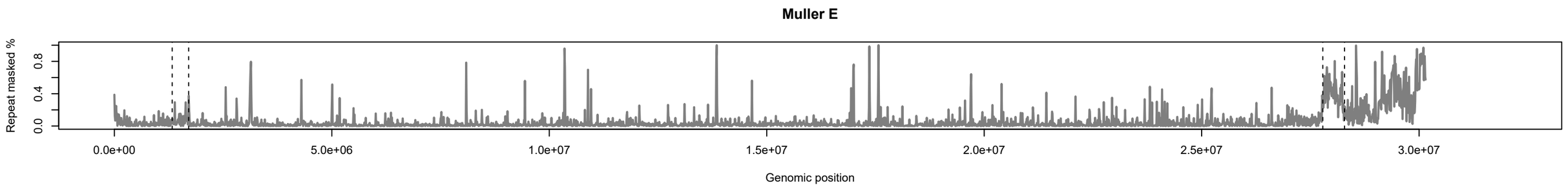

Muller A
Genes
0
2
4
6
8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

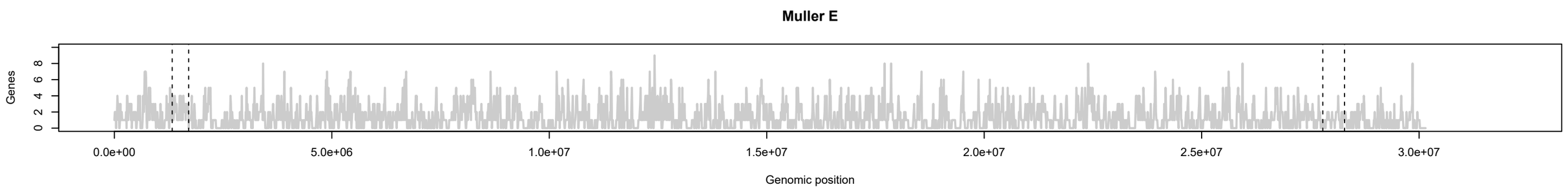

Muller B
GC %
0.2
0.4
0.6
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

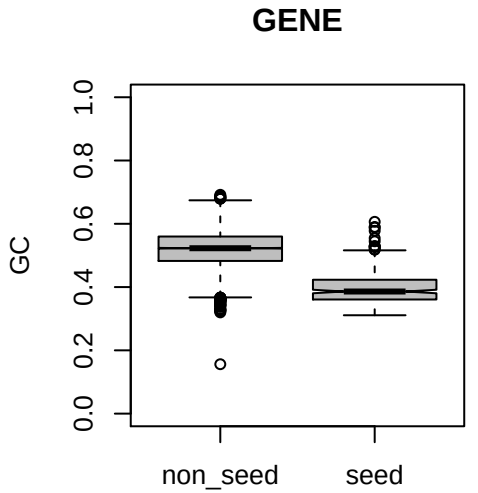

Muller B
Repeat masked %
0.0
0.4
0.8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

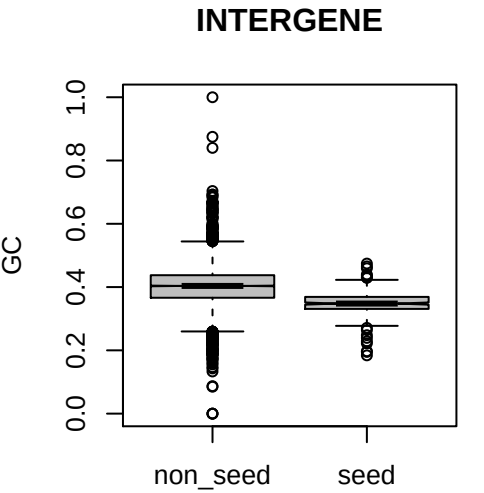

Muller B
Genes
0
2
4
6
8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

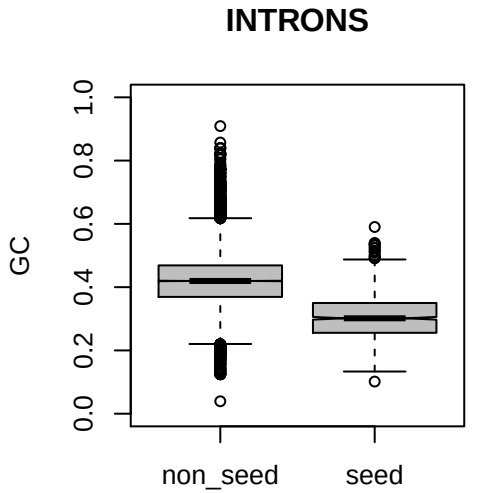

Muller E
GC %
0.2
0.4
0.6
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

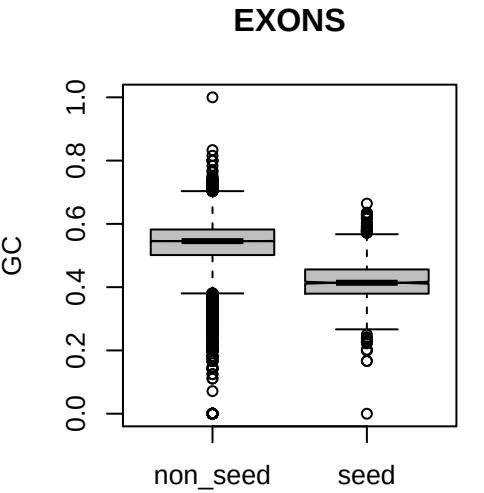

Muller E
Repeat masked %
0.0
0.4
0.8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

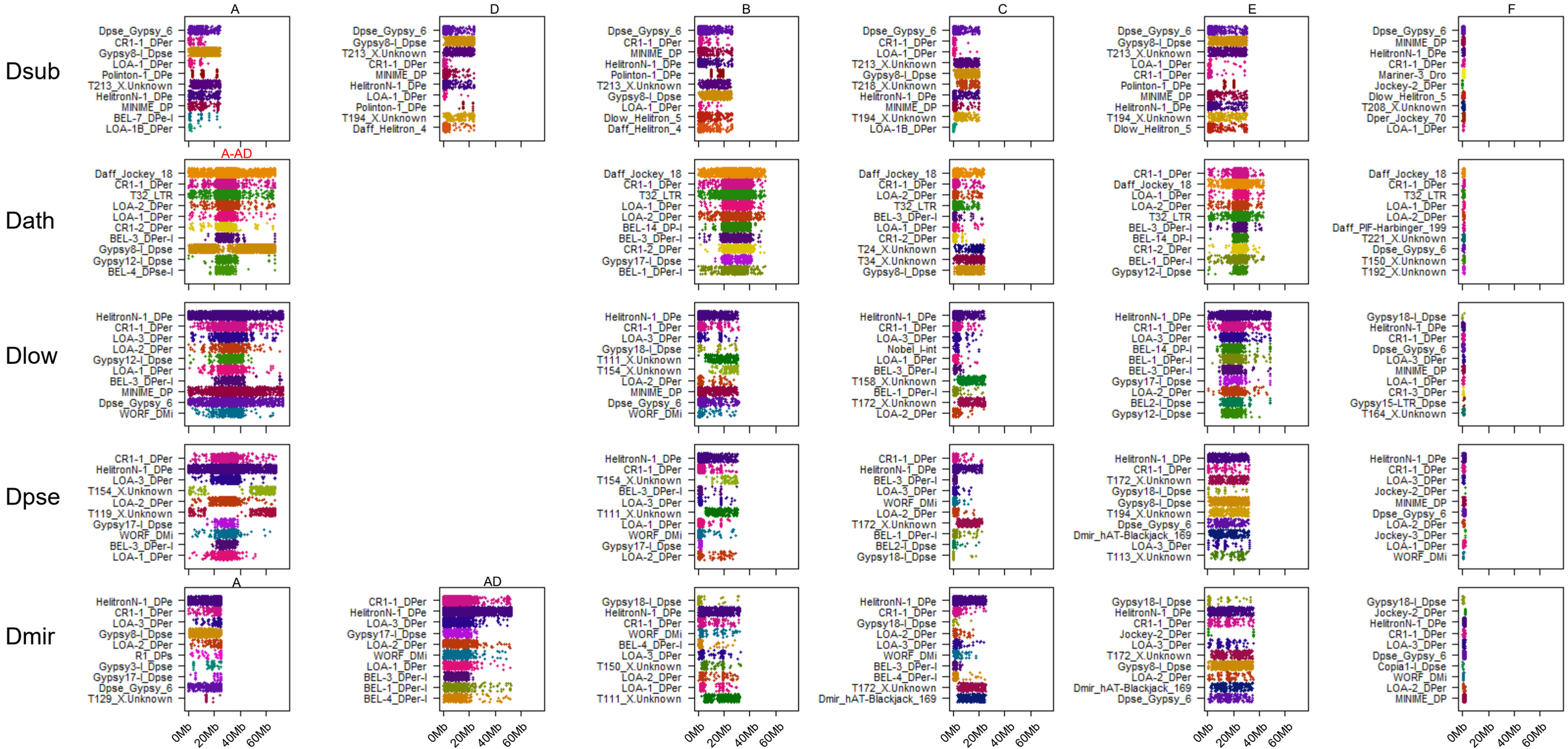

Muller E
Genes
0
2
4
6
8
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
Genomic position

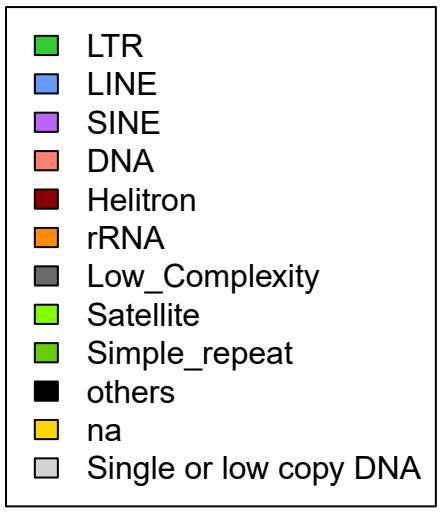

GENE
GC
0.0
0.2
0.4
0.6
0.8
1.0
non_seed
seed

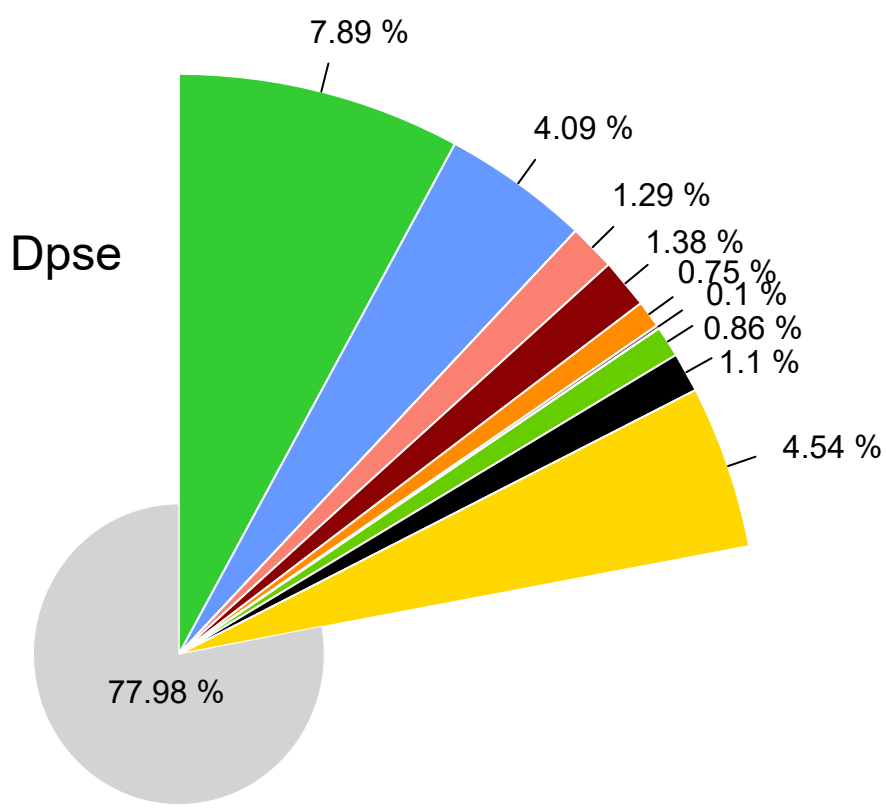

INTERGENE
GC
0.0
0.2
0.4
0.6
0.8
1.0
non_seed
seed

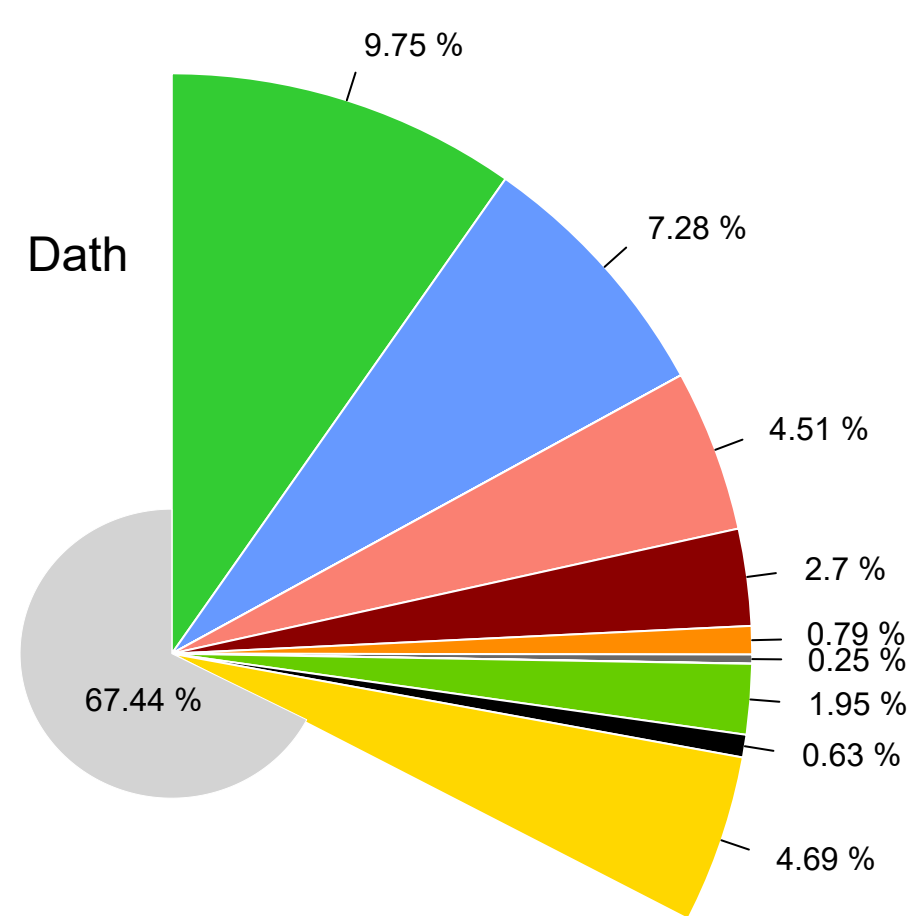

INTRONS
GC
0.0
0.2
0.4
0.6
0.8
1.0
non_seed
seed

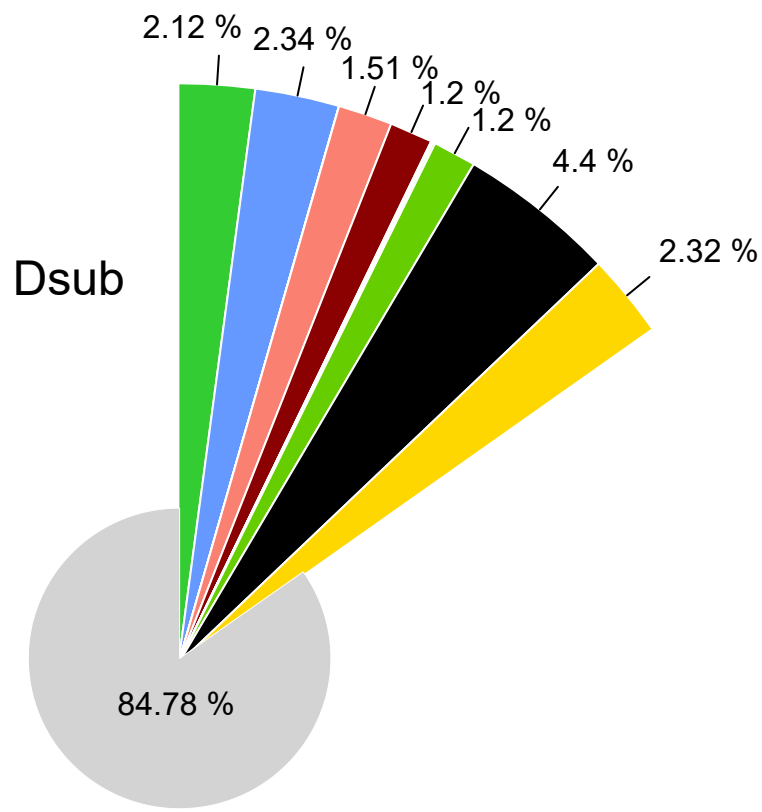

EXONS
GC
0.0
0.2
0.4
0.6
0.8
1.0
non_seed
seed

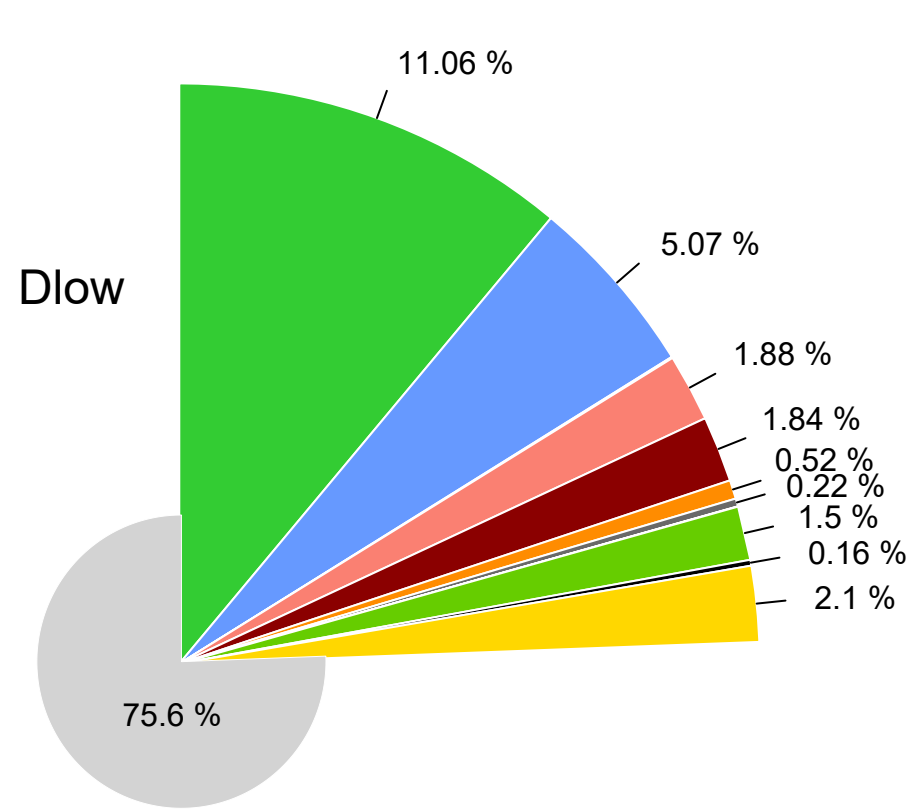

Dsub
Dath
Dlow
Dpse
Dmir
A
D
B
C
E
F
A-AD
AD
Dpse_Gypsy_6
CR1-1_DPer
Gypsy8-I_Dpse
LOA-1_DPer
Polinton-1_DPe
T213_X.Unknown
HelitronN-1_DPe
MINIME_DP
BEL-7_DPe-I
LOA-1B_DPer
T213_X.Unknown
Gypsy8-I_Dpse
CR1-1_DPer
MINIME_DP
HelitronN-1_DPe
LOA-1_DPer
Polinton-1_DPe
T194_X.Unknown
Daff_Helitron_4
Dlow_Helitron_5
T218_X.Unknown
T194_X.Unknown
Daff_Jockey_18
T32_LTR
LOA-2_DPer
BEL-3_DPer-I
BEL-14_DP-I
CR1-2_DPer
Gypsy17-I_Dpse
BEL-1_DPer-I
T24_X.Unknown
T34_X.Unknown
Daff_PIF-Harbinger_199
T221_X.Unknown
T150_X.Unknown
T192_X.Unknown
HelitronN-1_DPe
LOA-3_DPer
Gypsy18-I_Dpse
T111_X.Unknown
T154_X.Unknown
Nobel_I-int
T158_X.Unknown
T172_X.Unknown
BEL-14_DP-I
BEL-1_DPer-I
BEL-3_DPer-I
Gypsy17-I_Dpse
BEL2-I_Dpse
Gypsy12-I_Dpse
Gypsy15-LTR_Dpse
T164_X.Unknown
WORF_DMi
T154_X.Unknown
T119_X.Unknown
Gypsy18-I_Dpse
T172_X.Unknown
T194_X.Unknown
Dmir_hAT-Blackjack_169
T113_X.Unknown
Jockey-2_DPer
Jockey-3_DPer
R1_DPs
Gypsy3-I_Dpse
Gypsy17-I_Dpse
Dpse_Gypsy_6
T129_X.Unknown
BEL-4_DPer-I
T150_X.Unknown
T111_X.Unknown
T172_X.Unknown
Dmir_hAT-Blackjack_169
Gypsy18-I_Dpse
Jockey-2_DPer
Dpse_Gypsy_6
Copia1-I_Dpse
0Mb
20Mb
40Mb
60Mb


LTR
LINE
SINE
DNA
Helitron
rRNA
Low_Complexity
Satellite
Simple_repeat
others
na
Single or low copy DNA


Dpse
7.89 %
4.09 %
1.29 %
1.38 %
0.75 %
0.1 %
0.86 %
1.1 %
4.54 %
77.98 %


Dath
9.75 %
7.28 %
4.51 %
2.7 %
0.79 %
0.25 %
1.95 %
0.63 %
4.69 %
67.44 %


Dsub
2.12 %
2.34 %
1.51 %
1.2 %
1.2 %
4.4 %
2.32 %
84.78 %


Dlow
11.06 %
5.07 %
1.88 %
1.84 %
0.52 %
0.22 %
1.5 %
0.16 %
2.1 %
75.6 %

Appendix 1

For each Drosophila genome, we followed a similar assembly approach to (Mahajan *et al.* 2018). Briefly, long-read sequencing data were used to generate sequence contigs, Illumina data were used to polish the contigs and identify non-target contigs (contamination), and Hi-C data were used to place contigs into chromosomal scaffolds. However, given differences in the long-read datatype (PacBio or Nanopore), the amount of polymorphism present in the strains sequenced, the sex of the individuals used for sequencing, and differences in the genome architecture (such as repeat content) across species, slightly different species-specific adjustments were done during each assembly, to maximize the quality of each Drosophila genome. What follows is a description and characterization of steps taken during the assembly process for each species. In all cases, genome assemblies were initially interrogated using a variety of metrics to help determine the assembly quality and identify sex chromosome contigs and contaminant sequence (see METHODS). We used Bandage (Wick *et al.* 2015) to visualize our preliminary assemblies and metrics of interest were displayed overtop the assembly (**Appendix figure 1**). By using Bandage, we were able to easily explore various assembly algorithm parameters and their impacts on the quality of the draft assembly. Further, we could filter the assemblies in an efficient manner and find additional contig associations in the graph that were further validated using Hi-C associations, and in some cases, contigs could be merged when appropriate (see below).

Appendix figure 1. Bandage plot of a typical Drosophila genome assembly. The left panel is a visualization of the genome graph (.gfa file) from a canu assembly with the node name for each contig and Illumina coverage displayed in text overtop each contig. Each contig in the assembly is shaded by the amount of male Illumina whole genome sequencing coverage (see Methods). In this example, red contigs are likely autosomal (~40×) while darker contigs have less coverage and indicate either putative sex chromosome contigs (~20×) or putative contaminant contigs (<< 20×). B) Shown is a zoomed in image of 2 nodes (535 and 511) in the assembly with exceptionally low male (top) and female (bottom) Illumina coverage (< 0.1×). By also visualizing the top BLAST hits for these contigs (not shown), we were able to identify these contigs as belonging to an *Acetobacter* species and were thus contaminants marked for removal from the assembly. Contigs with exceptionally high Illumina coverage were also scrutinized thoroughly but these can arise for multiple reasons, including mtDNA contigs, collapsed regions of the target genome (e.g., rDNA genes or centromeric satellite sequence), or non-target contaminant contigs.

Drosophila miranda

The *D. miranda* genome was previously published, and the quality (both in terms of accuracy and contingency) of this assembly was confirmed with several orthogonal data sets including optical maps and extensive BAC clone sequencing (Mahajan *et al.* 2018). A detailed description of the assembly methods, including extensive QC at each assembly step are given in (Mahajan *et al.* 2018), and the assembly was also validated via statistical methods and short-read Illumina mapping (S3 Table in Mahajan *et al.* 2018). The quality and large-scale structural continuity of the *D. miranda* assembly was assessed by comparing it to sequenced BAC clones and optical mapping data. A total of 383 randomly selected BAC clones from a *D. miranda* male BAC clone library were shotgun sequenced (which should cover roughly 1/4 of the *D. miranda* genome), and 97% contiguously mapped to a unique position in the genome (see Fig 2D; S5 Table; S6 Table; S6 Fig in Mahajan *et al.* 2018). Similarly, most of the *D. miranda* genome was found to be covered by optical mapping data (S5 Fig and S4 Table in Mahajan *et al.* 2018). Thus, continuous and unique mapping of most BAC clones and coverage by optical reads confirmed the quality of the genome assembly.

Most importantly, BAC clone sequencing confirmed centromere assemblies in *D. miranda* (see S11 Fig from Mahajan *et al.* 2018), and BAC clone sequencing also confirm the assembly of 'paleocentromeres' in *D. miranda* (**Appendix figure 2**). Additionally, the presence of heterochromatic repeat island on Muller B and Muller E in *D. miranda* (that is, the paleocentromeres) were confirmed by FISH in polytene chromosomes of *D. miranda* (see Figure 1C of Mahajan *et al.* 2018). Thus, our comprehensive and detailed validation in *D. miranda* demonstrated that our approach is accurate in assembling heterochromatic regions, and similar approaches were used to assemble the genomes of the other species that we report here.

Appendix figure 2. BAC clone sequencing confirms centromere and pealeocentromere assembly. Several independent BAC clones map to the assembly of our paleocentromeres.

Drosophila subobscura

Our *D. subobscura* assembly was our most contiguous assembly (**Appendix table 1**) and had the fewest contigs in the initial canu/WTDBG2 assembly (**Appendix table 1**). This is probably due to the low density of repetitive elements in this genome (see RESULTS), and others have noted that *D. subobscura* and close relative *D. guanche* were easier to assemble and produced more contiguous assemblies than other Drosophila species (Puerma *et al.* 2018; Karageorgiou *et al.* 2019). The subsequent polishing steps of the genome assembly with Racon and Pilon (Walker *et al.* 2014; Vaser *et al.* 2017) improved the assembly and at each round the number of single copy BUSCOs (Simão *et al.* 2015) increased, while the number of missing and fragmented BUSCOs decreased (**Appendix table 1**). Hi-C scaffolding allowed us to orient contigs into chromosomes and all Muller elements were captured in < 10 contigs, and typically just a few contigs made up the bulk of the chromosome (See RESULTS). Multiple small contigs showed aberrant Illumina coverage patterns and BLAST to *Acetobacter spp.* and *Saccharomyces cerevisiae* (used as Drosophila food and found in the gut). This contaminating sequence (8 contigs amounting to 2.2 Mb) was removed from the assembly before final scaffolding (**Appendix table 1**). In our assembly, the number of contigs per chromosome was slightly lower, and the total length of each chromosome similar to a recently published PacBio-based assembly of the *ch-cu* line (Karageorgiou *et al.* 2019). Importantly, the centromeric seed regions on Muller A, B and E were all assembled in contigs in our assembly. Interestingly, the seed region on Muller B (Chr U) is also associated with a well-known inversion breakpoint that results in a segregating inversion that plays a role in environmental adaptation (Karageorgiou *et al.* 2019). Our assembly captures this inverted chromosome and provides a resource for further study of this inversion in *D. subobscura*.

Appendix table 1. Summary statistics and BUSCO results from the genome assembly process of *Drosophila subobscura.*

	canu/WTDBG2 only	**canu/WTDBG2 + Racon (3x)**	**canu/WTDBG2 + Racon (3x) + Pilon (1x)**	**canu/WTDBG2 + Racon (3x) + Pilon (2x)**	**Final Hi-C scaffolded Dsub_1.0**
N50	11,277,487	11,360,548	11,375,305	11,370,518	*
Max Contig	25,648,096	25,831,582	25,849,837	25,836,392	*
Assembly Size	128,396,075	129,296,034	129,434,338	129,376,892	126,232,139

Number of Contigs	62	61	61	61	*
Complete BUSCOs	987	1,017	1,057	1,060	1,062
Complete Single Copy BUSCOs	984	1,011	1,047	1,050	1,054
Duplicated	3	6	10	10	8
Fragmented	46	29	2	-	-
Missing	33	20	7	6	4
% BUSCOs complete	92.6%	95.4%	99.2%	99.4%	99.6%

Drosophila athabasca

For *D. athabasca*, we produced assemblies for two different lines belonging to different semispecies (EA and EB). Flow cytometry estimates of genome size were ~210Mb for both EA and EB.

EB assembly: For the initial canu assembly we used the setting genomeSize=210Mb. Using Bandage and BLAST we found 10.4 Mb of non-target sequence (primarily *Acetobacter* and *Sacchoramyces*) and this sequence was removed. Nodes that showed unambiguous associations in the canu graph and had Hi-C associations during preliminary analyses were merged in the assembly graph. We merged the nodes 2416 and 2418 (total 34,554,387 bp), 2438 and 2440 (total 23,432,006 bp), 2396, 2474, 2476, and 2400 (total 22,894,561 bp), 2409 and 2411 (total 9,116,924 bp), 58881, 58882, and 2424 (total 1,243,185 bp). We also split one contig (node 2401), which was a chimeric assembly that showed Illumina coverage patterns suggesting it being both autosomal and X-linked and which showed split autosomal/X-linked associations in our preliminary Hi-C analyses. For Muller C in the EB assembly, residual heterozygosity led to a fragmented assembly with many regions where fragments of both haplotypes were assembled. To try to improve contiguity for this Muller element and collapse heterozygous regions, we took all reads that mapped uniquely to contigs identified as belonging to Muller C and reassembled them using Minimap2/Miniasm (Li 2018), which in our experience often collapses heterozygous regions. We used the miniasm flags -s 1000 -d 500000 and inspected the resulting graph file (.gfa) and merged contigs in the graph with unambiguous connections. We merged nodes utg000002l, utg000034l, utg000039l, and utg000029l (total 3,929,811 bp), utg000010l and utg000036l (total 6,966,474 bp) and these plus the remaining contigs were used in place of the canu-based contigs of Muller C for final Hi-C scaffolding. Hi-C scaffolding clustered 22 Muller C contigs, but because of the shorter contig lengths and ambiguities in their placement due to the Hi-C being derived from the other semispecies, we assigned each contig to Muller C and named them Muller_C 1-22.

Our EB assembly statistics (**Appendix table 2**) improved at every step of the process and the number of complete BUSCOs increased with Quiver and Pilon genome polishing. Additional non-target sequence (total 3Mb) was identified and removed after our Quiver runs (**Appendix table 2**). The final scaffolded assembly contained long contigs that made up the bulk of the euchromatic arms for metacentric chromosomes (**Appendix figure 3**) and contigs extended into the pericentromere in all cases which aided in the assembly of the pericentromeric region. Due to the fragmented Muller C in our EB

assembly, we replaced the 22 Muller C EB contigs (above) with the scaffolded Muller C assembled from the EA (below) to create the 'Dath_EB_hybrid' genome (**Appendix table 2**). Muller C EB SNPs were replaced with EA SNPs (see METHODS). This assembly had the highest genome completeness score of any of the *D. athabasca* assemblies (**Appendix table 2**) and is what we present in the paper.

Appendix table 2. Summary statistics and BUSCO results from genome assembly process of *Drosophila athabasca EB*

	canu + bandage	***canu + Quiver (2x)***	***canu + Quiver (2x) + Pilon (2x)***	***Final Hi-C scaffolded Dath_EB_1.0***	***Final Hi-C scaffolded Dath_EB_hybrid***
N50	14,480,452	15,318,533	15,319,690	*	*
Max Contig	34,554,387	34,873,651	34,879,405	*	*
Assembly Size	195,713,124	192,740,469	192,655,157	192,660,667	192,054,219
Number of Contigs	199	133	133	*	*
Complete BUSCOs	1,043	1,054	1,057	1,057	1,060
Complete Single Copy BUSCOs	1,023	1,046	1,048	1,048	1,052
Duplicated	20	8	9	9	8
Fragmented	12	3	1	1	1
Missing	11	9	8	8	5
% BUSCOs complete	97.9%	98.9%	99.2%	99.2%	99.5%

Appendix figure 3. *Drosophila athabasca* EB metacentric chromosome Hi-C associations and scaffolding. Our EB assembly (**Appendix table 2**) was superior to our EA assembly (**Appendix table 3**) and long contigs from our EB assembly extended at least a megabase into the pericentromere for all metacentric chromosomes. Shown above are Hi-C association heatmaps from Juicebox (Durand *et al.* 2016). Green boxes denote contigs. The pericentromeric region is highlighted in purple, and note the clear transition in Hi-C associations between euchromatic and heterochromatic regions. We used EA Hi-C data to scaffold the EB assembly. The EA and EB semispecies harbor inversions that differentiate the semispecies and we identified numerous inversions when mapping EA Hi-C to the EB genome assembly. Thus, the exceptionally long EB contigs that extend into the pericentromeric region allowed us to accurately scaffold chromosomes while simultaneously identifying inversions along the euchromatic arms.

EA assembly: The initial canu assembly used the same parameters as used for the EB assembly (above and see METHODS). Our investigation of the resulting assembly using Bandage and BLAST identified 15.5 Mb of non-target sequence (primarily *Acetobacter* and *Sacchoramyces*) which was removed. Using Bandage we merged the unambiguous nodes 1334, 1335, 1336, 1337, 1338, 1340, 1342, and 1344 (total 11,326,223 bp), 1385 and 1387 (total 3,563,450 bp), 1369 and 162 (total 2,257,788 bp), 1328, 1330, and 1332 (total 5,082,211 bp). Assembly statistics for the EA assembly (**Appendix table 3**) did not approach those of the EB assembly, likely do to the shorter raw PacBio reads that we had for this genome assembly (see METHODS). BUCSO statistics improved during genome polishing (**Appendix table 3**) but were also slightly lower than our EB assembly (**Appendix table 2**). Hi-C scaffolding was able to assemble

the euchromatic arms and pericentromeric regions for all Muller elements (**Appendix figure 4**). However, for Muller E, we found many uncollapsed haplotypes due to residual heterozygosity in the sequenced line. In addition, the overall shorter contig lengths failed to assemble across the euchromatic/heterochromatic boundary for some Muller elements, leaving the orientation of the euchromatic arm and the pericentromeric region more ambiguous (**Appendix figure 4**).

Appendix table 3. Summary statistics and BUSCO results from genome assembly process of *Drosophila athabasca EA*

	canu + bandage	***canu + Quiver (2x) + Pilon (2x)***	***Final Hi-C scaffolded Dath_EA_1.0***
N50	5,537,664	5,538,275	*
Max Contig	20,031,442	20,052,631	*
Assembly Size	193,369,473	193,423,818	193,434,778
Number of Contigs	348	348	*
Complete BUSCOs	1,041	1,055	1,055
Complete Single Copy BUSCOs	1,021	1,041	1,041
Duplicated	20	14	14
Fragmented	14	5	5
Missing	11	6	6
% BUSCOs complete	97.7%	99.0%	99.0%

Appendix figure 4. *Drosophila athabasca* EA assembly Hi-C associations and scaffolding. A) Hi-C scaffolding of the EA assembly recovered long blocks of contigs we identified as Muller elements from our EB assembly. Blue boxes bound putative Muller element boundaries; green boxes denote contigs. Black arrows show contigs that span the euchromatic/heterochromatic transition and allow for confident scaffolding into pericentromeric regions. Blue arrows show regions where contigs failed to assemble across the transition making scaffolding based on Hi-C associations more challenging. B) Shown is a zoomed in image of Muller C scaffolding. Here, a contig spans the euchromatic/heterochromatic transition on Muller C and we used this scaffolded in our Dath_EB_hybrid assembly. Note the lack of evidence for inversions in the Hi-C heatmaps since here we are using EA Hi-C with an EA assembly.

Drosophila lowei

Drosophila lowei was the most fragmented of our assemblies (**Appendix table 4**), probably because we sequenced an outbred strain that harbored high levels of heterozygosity (*D. lowei* is extremely difficult to maintain in the lab). Excessive heterozygosity is known to be challenging for assembly algorithms. The subsequent polishing steps using Racon and Pilon improved the assembly by reducing both the number of fragmented genes and missing genes and increasing the number of complete BUSCOs (**Appendix table 4**). Our contig filtering process (see METHODS and above) identified multiple contigs with BLAST hits to *Acetobacter spp.* and *Saccharomyces cerevisiae* that also showed aberrant Illumina coverage patterns not consistent with our target species coverage (described above). These flagged contigs (amounting to a total of 7.1 Mb of sequence) were removed from the assembly. Although the final assembly was more fragmented when compared to our other assemblies, Hi-C scaffolding oriented contigs into scaffolds that appeared highly accurate. For example, Muller B was assembled into a scaffold that was remarkably similar to Muller B of *D. pseudoobscura* (see RESULTS) and the Muller C pericentromeric region appeared largely syntenic across all obscura group species, including *D. lowei* (See RESULTS). In total, our results provide some evidence that for even our most fragmented assembly,

we could correctly scaffold chromosomes using our Hi-C data. Hi-C scaffolding and contaminant removal during the final scaffolding step did lead to a small decrease in our BUSCO statistics (**Appendix table 4**). Whole genome comparisons with close relative *D. miranda* helped confirm our scaffolding (**Appendix figure** 5)

Appendix table 4. Summary statistics and BUSCO results from genome assembly process of *Drosophila lowei.*

	canu/WTDBG2 only	***canu/WTDBG2 + Racon (3x)***	***canu/WTDBG2 + Racon (3x) + Pilon (1x)***	***canu/WTDBG2 + Racon (3x) + Pilon (2x)***	***Final Hi-C scaffolded Dlow_1.0***
N50	4,754,630	4,787,044	4,797,192	4,793,318	*
Max Contig	20,816,730	20,907,879	20,946,520	20,929,419	*
Assembly Size	192,748,718	191,586,436	191,816,863	191,620,915	184,313,494
Number of Contigs	943	726	726	726	*
Complete BUSCOs	936	975	1,037	1,039	1,036
Complete Single Copy BUSCOs	920	960	1,016	1,021	1,022
Duplicated	16	15	21	18	14
Fragmented	67	39	3	1	1
Missing	63	52	26	26	29
% BUSCOs complete	87.8%	91.5%	97.3%	97.5%	97.2%

Appendix figure 5. Whole genome alignment of our assembly (Y axis) to the published *Drosophila miranda* genome.

Drosophila pseudoobscura

For *Drosophila pseudoobscua* we were interested in assembling the Muller elements and as much of the Y chromosome as possible. Therefore, we only Nanopore sequenced males. The availability of female Nanopore data of the same line (Miller *et al.* 2018) allowed us to increase our coverage for the autosomes and X chromosome by merging datasets. However, the varied coverage over each genomic partition (Autosomes > X > Y) led us to take an approach where we assembled each separately and then used Hi-C to scaffold the contigs together. We first targeted the autosomes for assembly using canu (see METHODS) and genomeSize=200m. All the resulting contigs > 0.8 Mb that mapped to Muller elements B, C, E and F in *D. miranda* using MUMmer (Kurtz *et al.* 2004) were marked as autosomal. We then polished the whole assembly (Racon and Pilon) and then mapped the raw Nanopore reads to the assembly and subtracted out uniquely mapping autosomal reads. The remaining reads were deemed enriched for X or Y and assembled using canu (see METHODS) targeting the X chromosome with genomeSize=100m. We followed methods outlined above (i.e. mapping of resulting contigs to Muller A-AD from *D. miranda*) to assemble and then polished the assembly. Finally, using all Nanopore reads, we used the same subtraction method (above) and produced reads enriched for Y and unincorporated

heterochromatin and performed a canu assembly with genomeSize=50m We then polished this final assembly using rounds of Racon and Pilon (**Appendix Table 5**). To finish, we again used Hi-C scaffolding to orient contigs from the entire assembly (autosomes, X and Y) while noting each contigs initial placement (above) and checking expected female/male Illumina coverage patterns for the autosomes, X chromosome, and Y to confirm assembly. Y chromosome contigs were not scaffolded.

We compared our genome assembly with the Dpse_r3.04 and found widespread concordance (**Appendix figure 6**). However, we see an improvement over the published assembly with regards to Muller B and Muller A-AD and there is a substantial increase in the amount of pericentromeric sequence assigned to all Muller elements in our assembly (**Appendix figure 6**). BUSCO statistics were found to be rather high and comparable to our other genome assemblies.

Appendix table 5. Summary statistics and BUSCO results from the genome assembly process of *Drosophila pseudoobscura*

	canu + Racon (3x)	**canu + Racon (3x) + Pilon (1x)**	**canu + Racon (3x) + Pilon (2x)**	**Final Hi-C scaffolded Dpse_1.0**
N50	5,996,964	5,975,358	5,971,646	*
Max Contig	20,387,169	20,334,965	20,319,488	*
Assembly Size	194,744,540	194,172,171	193,935,436	193,980,066
Number of Contigs	481	481	481	*
Complete BUSCOs	974	1,052	1,057	1,055
Complete Single Copy BUSCOs	968	1,043	1,048	1,048
Duplicated	6	9	9	7
Fragmented	59	3	2	3
Missing	33	11	7	8
% BUSCOs complete	91.4%	98.6%	99.2%	99.0%

Appendix figure 6. Whole genome alignment of our assembly (Y axis) to the published *Drosophila pseudoobscura* genome assembly (version 3.04). Scaffolds in the published assembly that are near chromosome length (i.e., Muller E and Muller C) largely agree with our scaffolds. However, our assembly extends the assembled length of these chromosomes with far less scaffolding. For Muller B and Muller A-AD, our scaffolded chromosomes show large stretches of collinearity with the fragmented published assembly, with the exception of a few inverted regions. Our Hi-C data and association heatmap (see RESULTS) argue that our assembly orientation is likely the correct one and provide orientation to the 5 large scaffolds of Muller B and 8 scaffolds of Muller A-AD in the reference genome. The paleocentromeric region on Muller E is assembled in the current reference genome, but our assembly contains additional sequence not present in the published assembly (not shown).

Durand N. C., Robinson J. T., Shamim M. S., Machol I., Mesirov J. P., Lander E. S., Aiden E. L., 2016 Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst **3**: 99–101.

Karageorgiou C., Gámez-Visairas V., Tarrío R., Rodríguez-Trelles F., 2019 Long-read based assembly and synteny analysis of a reference Drosophila subobscura genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. BMC Genomics **20**: 223.

Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., Antonescu C., Salzberg S. L., 2004 Versatile and open software for comparing large genomes. Genome Biol. **5**: R12.

Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences. (I Birol, Ed.). Bioinformatics **34**: 3094–3100.

Mahajan S., Wei K. H.-C., Nalley M. J., Gibilisco L., Bachtrog D., 2018 De novo assembly of a young Drosophila Y

chromosome using single-molecule sequencing and chromatin conformation capture. (C Tyler-Smith, Ed.). PLoS Biol. **16**: e2006348.

Miller D. E., Staber C., Zeitlinger J., Hawley R. S., 2018 Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3 (Bethesda) **8**: 3131–3141.

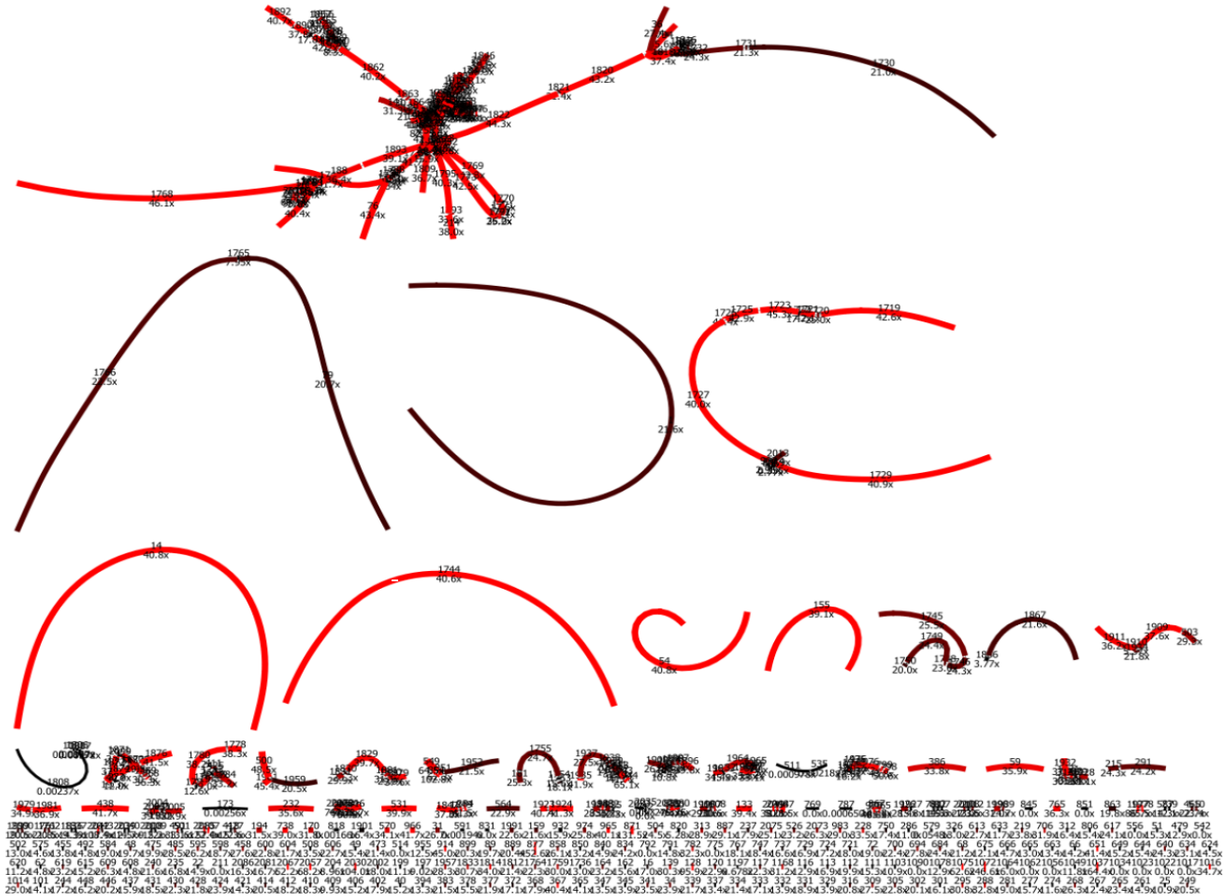
Puerma E., Orengo D. J., Cruz F., Gómez-Garrido J., Librado P., Salguero D., Papaceit M., Gut M., Segarra C., Alioto T. S., Aguade M., 2018 The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability. (J Gonzalez, Ed.). Genome Biol Evol **10**: 1956–1969.

Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M., 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**: 3210–3212.

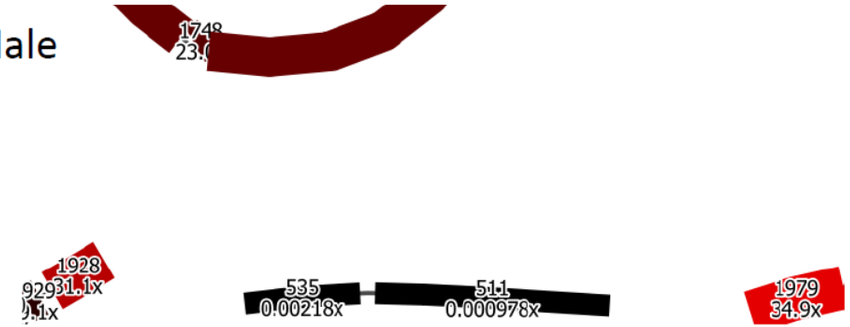
Vaser R., Sović I., Nagarajan N., Šikić M., 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. **27**: 737–746.

Walker B. J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C. A., Zeng Q., Wortman J., Young S. K., Earl A. M., 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. (J Wang, Ed.). PLoS ONE **9**: e112963.

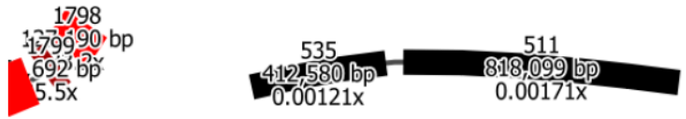
Wick R. R., Schultz M. B., Zobel J., Holt K. E., 2015 Bandage: interactive visualization of de novo genome assemblies. Bioinformatics **31**: 3350–3352.

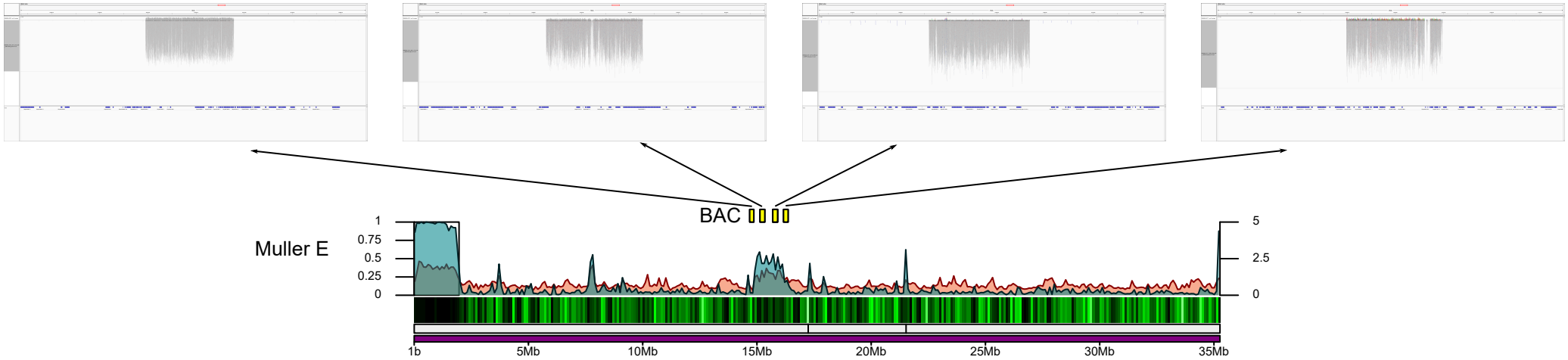


Male


1928
535
0.00218x
511
0.000978x
1979
34.9x

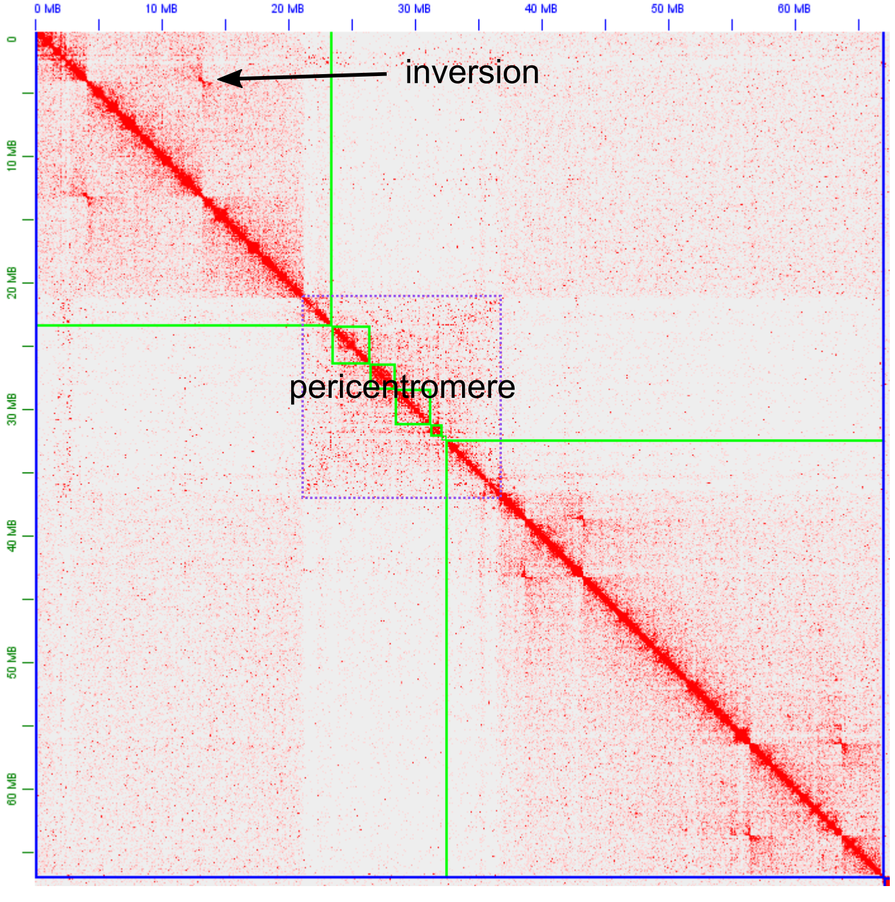
Female


1798
535
511
818,099 bp
0.00121x
0.00171x

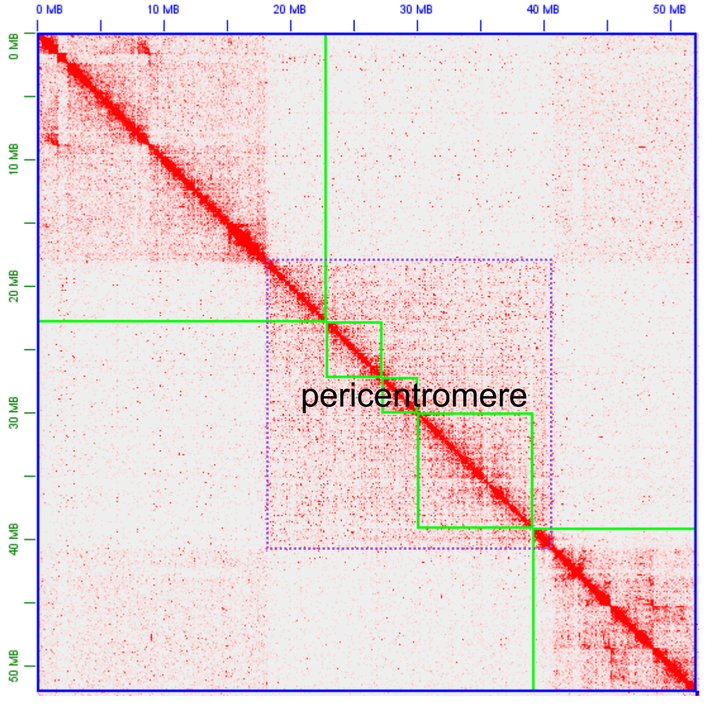

BAC
Muller E
1
0.75
0.5
0.25
0
5
2.5
0
1b
5Mb
10Mb
15Mb
20Mb
25Mb
30Mb
35Mb

Dath EB

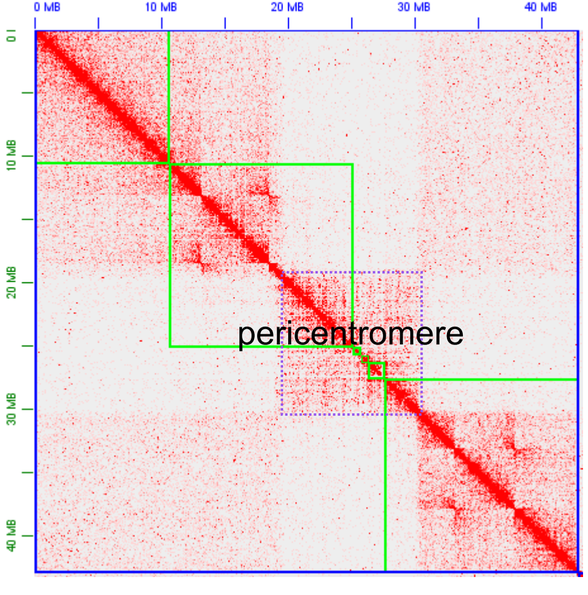
Muller A-AD

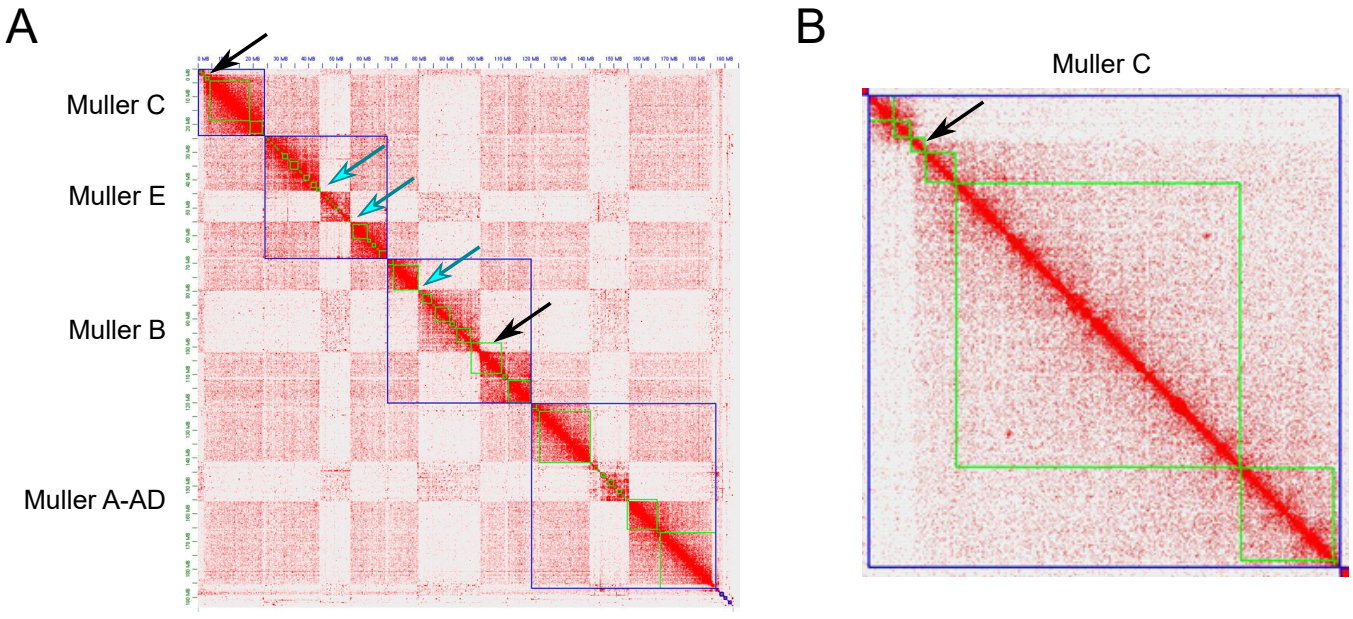


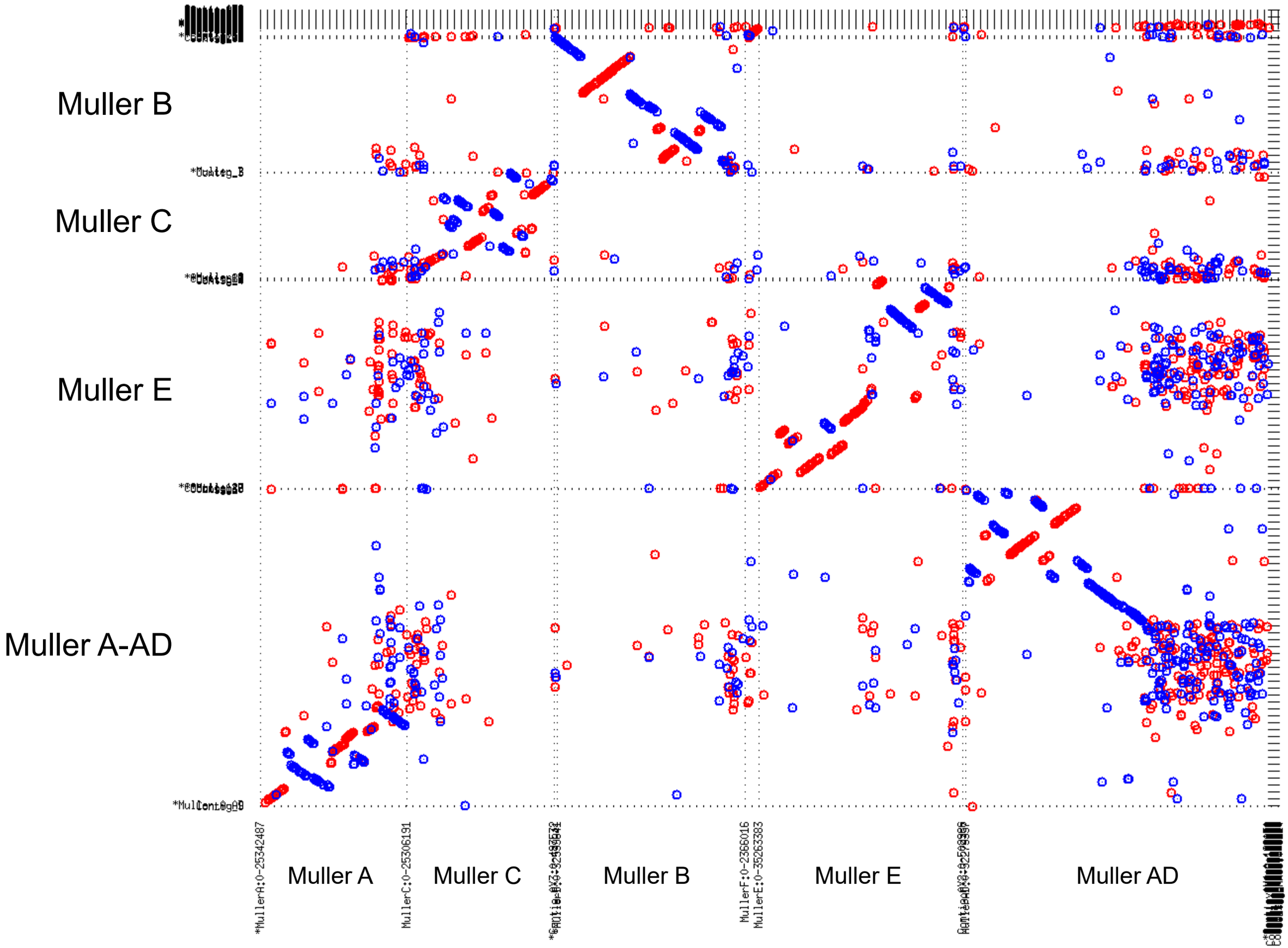
Muller B



Muller E




A
Muller C
Muller E
Muller B
Muller A-AD
B
Muller C


Muller B
Muller C
Muller E
Muller A-AD
Muller A
Muller C
Muller B
Muller E
Muller AD
Dmir

Muller C

Muller A-AD

Muller E

Muller B

Dpse_r3.04