PLOS BIOLOGY

METHODS AND RESOURCES

De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture

Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco, Doris Bachtrog*

Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America

* dbachtrog@berkeley.edu

OPEN ACCESS

Citation: Mahajan S, Wei KH-C, Nalley MJ, Gibilisco L, Bachtrog D (2018) De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture. PLoS Biol 16(7): e2006348. https://doi.org/10.1371/journal.pbio.2006348

Academic Editor: Chris Tyler-Smith, The Wellcome Trust Sanger Institute, United Kingdom of Great Britain and Northern Ireland

Received: April 7, 2018

Accepted: July 4, 2018

Published: July 30, 2018

Data Availability Statement: Experimental raw data are available from the NCBI SRA database. The genome sequence and annotation are available from NCBI. A full list of accession numbers is in S11 Table.

Funding: NIH (grant number R01GM076007, R01GM101255, R01GM093182). Received by DB. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

While short-read sequencing technology has resulted in a sharp increase in the number of species with genome assemblies, these assemblies are typically highly fragmented. Repeats pose the largest challenge for reference genome assembly, and pericentromeric regions and the repeat-rich Y chromosome are typically ignored from sequencing projects. Here, we assemble the genome of *Drosophila miranda* using long reads for contig formation, chromatin interaction maps for scaffolding and short reads, and optical mapping and bacterial artificial chromosome (BAC) clone sequencing for consensus validation. Our assembly recovers entire chromosomes and contains large fractions of repetitive DNA, including about 41.5 Mb of pericentromeric and telomeric regions, and >100 Mb of the recently formed highly repetitive neo-Y chromosome. While Y chromosome evolution is typically characterized by global sequence loss and shrinkage, the neo-Y increased in size by almost 3-fold because of the accumulation of repetitive sequences. Our high-quality assembly allows us to reconstruct the chromosomal events that have led to the unusual sex chromosome karyotype in *D. miranda*, including the independent de novo formation of a pair of sex chromosomes at two distinct time points, or the reversion of a former Y chromosome to an autosome.

Author summary

Y chromosomes determine the gender in many species, but their molecular investigation has been hampered by a lack of high-quality sequence assemblies. Here, we create a genome assembly of unprecedented quality and contiguity for the fruit fly *Drosophila miranda*, a model for Y chromosome research, which allows us to reconstruct the evolutionary events that create and dismantle sex chromosomes. Our assembly recovers entire chromosomes and notoriously difficult regions to assemble, including entire centromeres, large repetitive gene families embedded in heterochromatin, and more than 100 Mb of the highly repetitive and heterochromatic Y chromosome. We identify the putative

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BAC, bacterial artificial chromosome; cenH3, centromeric variant of histone H3; MY, million years; PG, proximity-guided; TRF, Tandem Repeat Finder; Y_{anc}, ancestral Y chromosome; rDNA, ribosomal DNA.

centromeric repeat in this species, which shows no sequence homology to centromere motifs of other *Drosophila* species. The recovered Y/neo-Y sequence is over three times the size of its former homolog, the neo-X, challenging a paradigm of sex chromosome evolution: rather than shrinking—the fate that is typically ascribed to Y chromosomes—we find that early Y evolution is instead characterized by a global DNA gain. The ancestral Y chromosome of *Drosophila*, by contrast, has become linked to an autosome in *D. miranda*, and we reconstruct the genomic and epigenetic changes that likely occurred to revert this former Y to an autosome.

Introduction

Sex chromosomes are derived from ordinary autosomes, yet old X and Y chromosomes contain a vastly different gene repertoire [1]. In particular, X chromosomes resemble the autosome from which they were derived, with only few changes to their gene content [2]. In contrast, Y chromosomes dramatically remodel their genomic architecture. Y evolution is characterized by massive gene decay, with the vast majority of the genes originally present on the Y disappearing, and Y degeneration is often accompanied by the acquisition of repetitive DNA [3]; old Y chromosomes typically have shrunk dramatically in size and contain only few unique genes but vast amounts of repeats.

The decrease in sequencing cost and increased sophistication of assembly algorithms for short-read platforms have resulted in a sharp increase in the number of species with genome assemblies. Indeed, X chromosomes have been characterized and sequenced in many species. However, assemblies based on short-read technology are highly fragmented, with many gaps, ambiguities, and errors remaining; this is especially true for repeat-rich regions, such as centromeres, telomeres, or the Y chromosome [4–6]. Thus, most sequencing projects have ignored the Y chromosome. Labor-intensive sequencing of Y chromosomes in a few mammal species has revealed a surprisingly dynamic history of Y chromosome evolution, with meiotic conflicts driving gene acquisition on the mouse Y chromosome [7], or gene conversion within palindromes retarding Y degeneration in primates [8]. However, all current Y assemblies are based on tedious resequencing of bacterial artificial chromosome (BAC) clones and available only for a handful of species [9–11], and the repeat-rich nature of Y chromosomes has hampered their evolutionary studies in most organisms.

Here, we present a near-finished reference genome for *Drosophila miranda*, including its Y chromosome, using a combination of long-read single-molecule sequencing, high-fidelity short-read sequencing, optical mapping, BAC clones sequencing, and Hi-C-based chromatin interaction maps. *D. miranda* has become a model system for studying the molecular and evolutionary processes driving sex chromosome differentiation, because of its recently evolved neo-sex chromosome system (see **Fig 1**). In particular, chromosomal fusions within *D. miranda* have resulted in the recent sex-linkage of former autosomes at two independent time points (**Fig 1D**), and these new sex chromosomes are at different stages in their transition to differentiated sex chromosomes. Specifically, chromosomes XR and YD became sex-linked about 15 million years (MY) ago [12], and the neo-X and neo-Y became sex chromosomes only about 1.5 MY ago [13]. These former autosomes are in the process of evolving the stereotypical properties of ancestral sex chromosomes [14,15]. Intriguingly, the ancestral Y chromosome (Y_{anc}) in this species group became fused to an autosome, probably around the same time XR and YD formed, and lost some of the characteristics of an ancient Y chromosome

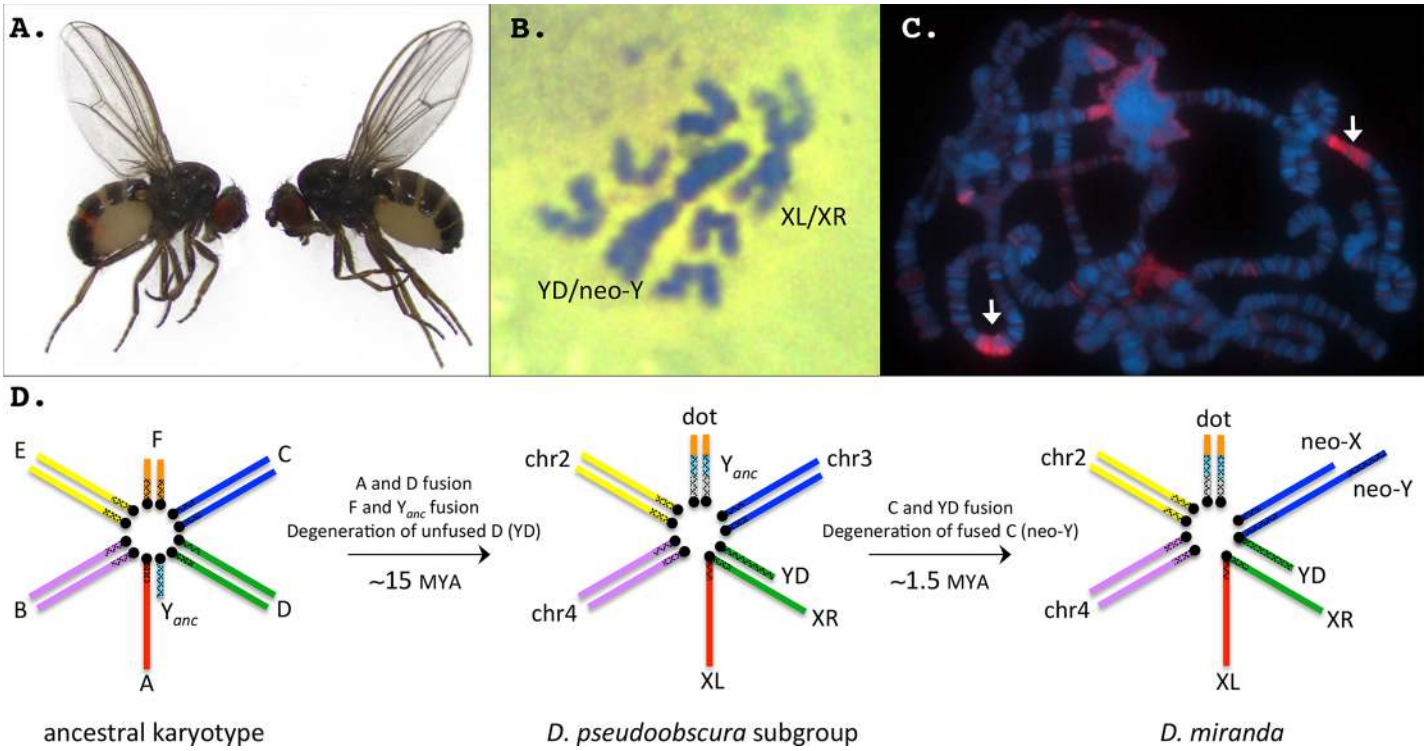

Fig 1. *Drosophila miranda* is a model species to study sex chromosome evolution. A. Male (left) and female (right) *D. miranda*. **B.** Mitotic chromosome squashes of male *D. miranda*. Both the ancestral X (XL/XR) and the Y chromosome (YD/neo-Y) show large blocks of dark staining (Giemsa), indicative of heterochromatin. The acrocentric rods are the neo-X, and chromosomes 2 and 4. **C.** Polytene chromosomes of a female *D. miranda* stained for *HP1* (heterochromatin protein 1). Note the large blocks of heterochromatin (arrows) on chromosomes 2 and 4. **D.** Karyotype evolution in *D. miranda*. Chromosomal fusions between the sex chromosomes and autosomes have resulted in both the reversal of Y_{anc} to an autosome as well as the independent de novo formation of new sex chromosomes from autosomes at two distinct evolutionary time points (XR and YD were formed about 15 MY ago, and the neo-X and neo-Y originated about 1.5 MY ago). Genome analysis allows us to reconstruct the temporal dynamics and molecular processes involved in sex chromosome evolution in this species. chr, chromosome; dot, dot chromosome; XL and XR, left and right arm of the X chromosome; YD, Y chromosome resulting from the unfused D element; *HP1*, heterochromatin protein 1; MY, million years; MYA, million years ago; Y_{anc}, ancestral Y chromosome.

https://doi.org/10.1371/journal.pbio.2006348.g001

[16–18]. Thus, *D. miranda* allows the investigation of the functional and evolutionary changes occurring on differentiating sex chromosomes, and their reversal.

The most recent assembly of *D. miranda* was generated via short-read Illumina sequencing and is highly fragmented. In particular, the genome was in 47,035 scaffolds, with a scaffold N50 (a weighted median statistic such that 50% of the entire assembly is contained in scaffolds equal to or larger than this value) of 5,007 bp and a total assembled genome size of 112 Mb (a female-only assembly resulted in 22,259 scaffolds, with an N50 of 13,773 bp and an assembled size of 125 Mb). The high amount of sequence similarity between the neo-sex chromosomes (98.5% identical at the nucleotide level), yet high repeat content of the neo-Y (over 50% of its DNA is derived from repeats [19]) posed a particular challenge to its assembly using short reads. Specifically, initial attempts to assemble the neo-Y resulted in a chimeric, highly fragmented and incomplete assembly, consisting of 36,282 (often chimeric) scaffolds, and a scaffold N50 of only 715 bp [20]. Thus, our previous analysis of neo-Y chromosome gene content evolution was instead based on mapping male reads to the neo-X assembly and identifying male-specific SNPs [20], or trying to reconstruct neo-Y transcripts using both male and female genome and transcriptome data [21]. This indirect approach, however, only allows the investigation of conserved regions on the neo-sex chromosome that differ by simple SNPs or short indels within genes. Here, we assemble the genome of *D. miranda* using long reads for contig formation, short reads

for consensus validation, and scaffolding by chromatin interaction mapping, and we verify our assembly using optical maps and BAC clone sequencing. Our assembly covers large fractions of repetitive DNA, with entire chromosomes being in a single scaffold, including their centromeres, and we recover over >100 Mb of the recently formed neo-Y chromosome. Our new assembly strategy achieves superior continuity and accuracy and provides a new standard reference for the investigation of repetitive sequences and Y chromosome evolution in this species.

Results and discussion

De novo assembly of a *D. miranda* reference genome

We sequenced adult male *D. miranda* (from the inbred strain MSH22) using a combination of different technologies: single-molecule real-time sequencing (PacBio), paired-end short-read sequencing (Illumina HiSeq), optical mapping (using BioNano), shotgun BAC clones sequencing (Illumina HiSeq) and chromatin conformation capture (Hi-C; see **S1 Table**).

Assembly of these complementary data types proceeded in a stepwise fashion (**Fig 2A**), similar to a recent approach [22], to produce progressively improved assemblies (**Table 1**). Briefly, we produced two initial assemblies of the PacBio data alone using the Falcon [23] and Canu [24] assemblers, and double merged the resulting assemblies with Quickmerge [25]. The resulting hybrid assembly had a contig NG50 (the minimum length of contigs accounting for half of the haploid genome size) of 5.2 Mb in 271 scaffolds. PacBio contigs were separated into X-linked and autosomal contigs versus Y-linked contigs, based on genomic coverage patterns of mapped male and female Illumina reads (**S1 Fig**), to avoid cross-mapping of short-read Hi-C data, and clustered into chromosome-scale scaffolds using Hi-C data (**Fig 2B**, **S2 Fig**, **S3 Fig**). Mapping of Illumina reads also allowed us to identify and remove contigs that resulted from uncollapsed haplotypes (**S4 Fig**). X-linked and autosomal contigs were scaffolded with female Hi-C libraries, while Y-linked contigs were clustered using Y-mapping reads from male Hi-C libraries (**S2 Fig**). Visual inspection of contact probability maps allowed us to identify a few misassemblies, which were manually corrected followed by rescaffolding (**S2 Table**). To assess quality, the resulting assembly was validated via statistical methods and short-read Illumina mapping (**S3 Table**), and comparison to optical mapping data (**S4 Table** and **S5 Fig**) and sequenced BAC clones from the MSH22 strain (**S5 Table**, **S6 Table**, **Fig 2D** and **S6 Fig**) and previous assemblies (*D. miranda* D.mir1.0 [20] **Fig 2C** and **S7 Fig**; *D. pseudoobscura;* **S8 Fig**).

To maximize accuracy of the final reference assembly, errors were manually curated before final gap filling and polishing (**S2 Table**). Our final assembly, D.mir2.0, totaled 287 Mb of sequence, with a scaffold NG50 of 35.3 Mb (**Table 1**). D.mir2.0 comprises just 102 scaffolds and 120 gaps (**S7 Table**), and the three autosomes, the three X chromosome arms, and the Y of *D. miranda* are all mostly covered by a single scaffold (**Fig 3**). The unplaced scaffolds are relatively small (median size 37.3 kb) and highly repeat-rich (median repeat content 94.7%), and sex-specific coverage patterns suggest that most are derived from the Y chromosome. In contrast, the previous assembly D.mir1.0 consisted of 47,035 scaffolds [20]. We used two approaches, REPdenovo [26] and RepeatModeler [27], to annotate repeats in the *D. miranda* genome and Maker [28] to annotate genes (**Fig 3**). We identified a total of 17,745 genes, and 43.7% of the genome was annotated as repeats. BUSCO assessments [29] support that our genome assembly and annotation are highly complete (**S8 Table**).

Assembly benchmarking and comparison to reference

The previous *D. miranda* reference assembly (D.mir1.0) was generated from paired-end short reads using the SOAPdenovo assembler and cross-species scaffold alignments to *D. pseudoobscura* [20]. Paired-end read sequences used to create the D.mir1.0 reference assembly

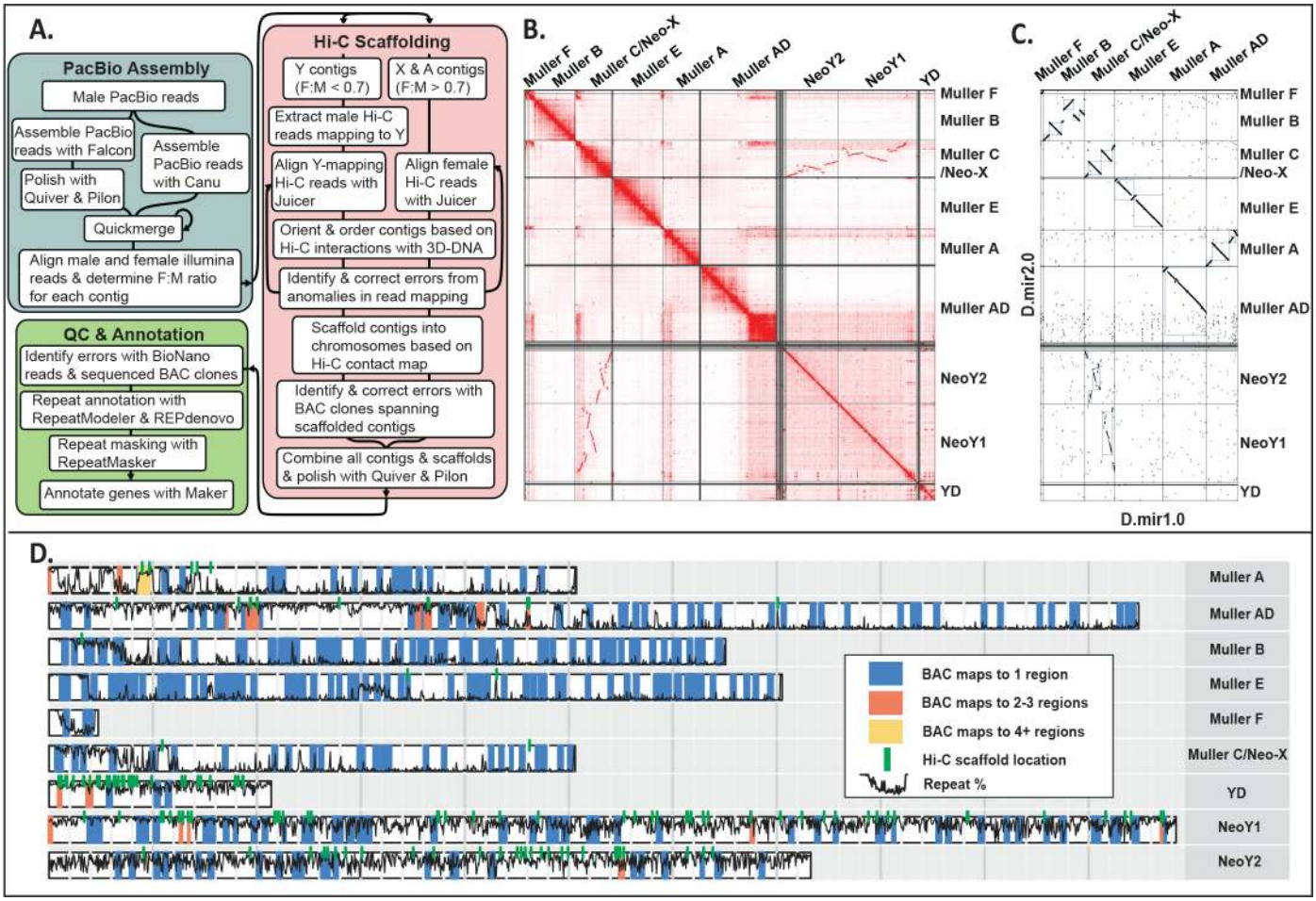

Fig 2. Assembly and validation of *Drosophila miranda* genome. A. Overview of assembly pipeline. The steps include assembly of male PacBio reads followed by scaffolding using Hi-C, and extensive QC using BioNano reads and BAC clone sequencing followed by gene and repeat annotation. **B.** Hi-C linkage density map. Chromatin interaction maps allow recovery of entire chromosome arms. Note that the Y-linked contigs were scaffolded separately from X-linked and autosomal contigs. Unlinked regions with many contacts indicate repetitive regions. **C.** Comparison of current (Dmir2.0) versus old (Dmir1.0) *D. miranda* assembly. Note that the Y/neo-Y was not assembled in Dmir1.0, and the dot plot indicates homology between our neo-Y assembly and the neo-X. Other repeat-rich regions, such as the large pericentromeric block on AD, are also missing from D.mir1.0. **D.** BAC clone mapping for assembly verification. BAC clones are color coded according to how many genomic regions they map to in our assembly; green lines indicate stitch points of scaffolds based on Hi-C contacts, and the black line gives the local repeat content along the genome. Three hundred sixty-one sequenced BAC clones (97%) map contiguously and uniquely to our genome assembly. BAC, bacterial artificial chromosome; F, female; M, male; QC, quality control; Repeat %, local repeat content.

https://doi.org/10.1371/journal.pbio.2006348.g002

were aligned to our D.mir2.0 assembly for a reference-free measure of structural correctness. These alignments confirmed that our current assembly is a dramatic improvement over D. mir1.0 (S3 Table), with fewer putative translocations (36 versus 17,764), deletions (229 versus 6,075), and duplications (8 versus 1,703). *D. miranda* and *D. pseudoobscura* are known to harbor dozens of inversions [30], and the initial *D. miranda* genome was scaffolded using *D. pseudoobscura*. Genome-wide alignments between our current *D. miranda* assembly and D.mir1.0 reveal dozens of rearrangements that were likely introduced by the scaffolding (Fig 2C; S7 Fig) and reflect inversions between *D. miranda* and *D. pseudoobscura* (see S7 Fig and S8 Fig).

We independently assessed the quality and large-scale structural continuity of our assembly by comparing it to sequenced BAC clones and optical mapping data. In total, we shotgun sequenced 383 randomly selected BAC clones from a *D. miranda* male BAC clone library [19], which should cover roughly 1/4 of the *D. miranda* genome. Three hundred seventy-two BAC

Table 1. Assembly statistics.

Assembly	Contigs + Scaffolds	Scaffolds	Unplaced Contigs	N50 (bp)	Assembly Size (Mb)	Assembly in Scaffolds (%)
PacBio Falcon	625	NA	625	2,242,328	273	NA
PacBio Canu	521	NA	521	3,884,273	296	NA
Quickmerge	271	NA	271	5,177,776	295	NA
PacBio + Hi-C	102	14	88	37,186,217	289	96.5
D.mir1.0 (female only, stitched with *Drosophila pseudoobscura*)	4,236	6	530	28,826,359	140	97.9
D.mir1.0 (not stitched with *D. pseudoobscura*)	47,035	NA	NA	5,007	112	NA
D.mir2.0; X-linked and autosomal scaffolds	40	6	34	32,539,883	177	97.1
D.mir2.0; Y-linked scaffolds	62	8	54	36,637,378	111	95.7
D.mir2.0	102	14	88	35,263,102	287	96.6

Abbreviations: NA, not applicable; N50, 50% of the assembly is contained in contigs or scaffolds equal to or larger than this value.

https://doi.org/10.1371/journal.pbio.2006348.t001

clones passed our sequence coverage filter and could be aligned to our *D. miranda* genome; of those, 361 (i.e., 97%) contiguously map to a unique position in the genome (Fig 2D; S5 Table, S6 Table; S6 Fig). Only 11 BAC clones map to two or more (typically highly repetitive) genomic locations (Fig 2D), and could represent assembly mistakes or recombinant BAC clones. Similarly, most of our genome is covered by optical mapping data (S5 Fig and S4 Table). Thus, continuous and unique mapping of most BAC clones and coverage by optical reads confirm the high quality of our genome assembly.

Assembly of highly repeat-rich regions

Our high-quality assembly contains large amounts of highly repetitive regions, including telomeres, pericentromeric regions, and putative centromeric repeats as well as the repeat-rich Y chromosome. Overall, about 126 Mb of the assembled 287 Mb *D. miranda* genome are repetitive, and we assembled about 41 Mb of pericentromeric and centromeric repeats and telomeres (S7 Table). In some cases, we assembled through the entire centromere and recovered telomeric repeats at the end of a chromosome arm (see below). In contrast, the previous *D. miranda* assembly based on only Illumina reads recovered less than 0.5 Mb of pericentromeric DNA (S7 Table, S7 Fig), and even the highly curated *D. melanogaster* genome assembly [31] entirely lacks centromeric sequence (S9 Fig). In addition, we assembled 110.5 Mb of Y-linked sequence, with 101.5 Mb contained within a single scaffold (Fig 3). Our assembly allows us to recover repetitive regions, including gene duplications and tandem repeats, most of which were collapsed and missed in our previous assembly (S10 Fig).

Recovery of chromosome ends and identification of putative centromeric DNA sequences

In *Drosophila*, telomeres are maintained by the occasional transposition of specific non-LTR retrotransposons (i.e., the HeT-A, TAHRE, and TART elements) to chromosome ends [32,33], and hybridization studies have suggested about two telomere repeats per chromosome end in *D. miranda* [34]. Indeed, for almost all chromosome arms (Muller A, B, C, both ends of E, F, neo-Y, and YD), we properly identified the ends of chromosomes based on the presence of telomeric transposable elements (see Fig 4A, Fig 4B).

Centromere sequences show little conservation between closely related species but have a common organization in most animals and plants [35–37]. In particular, centromeres typically

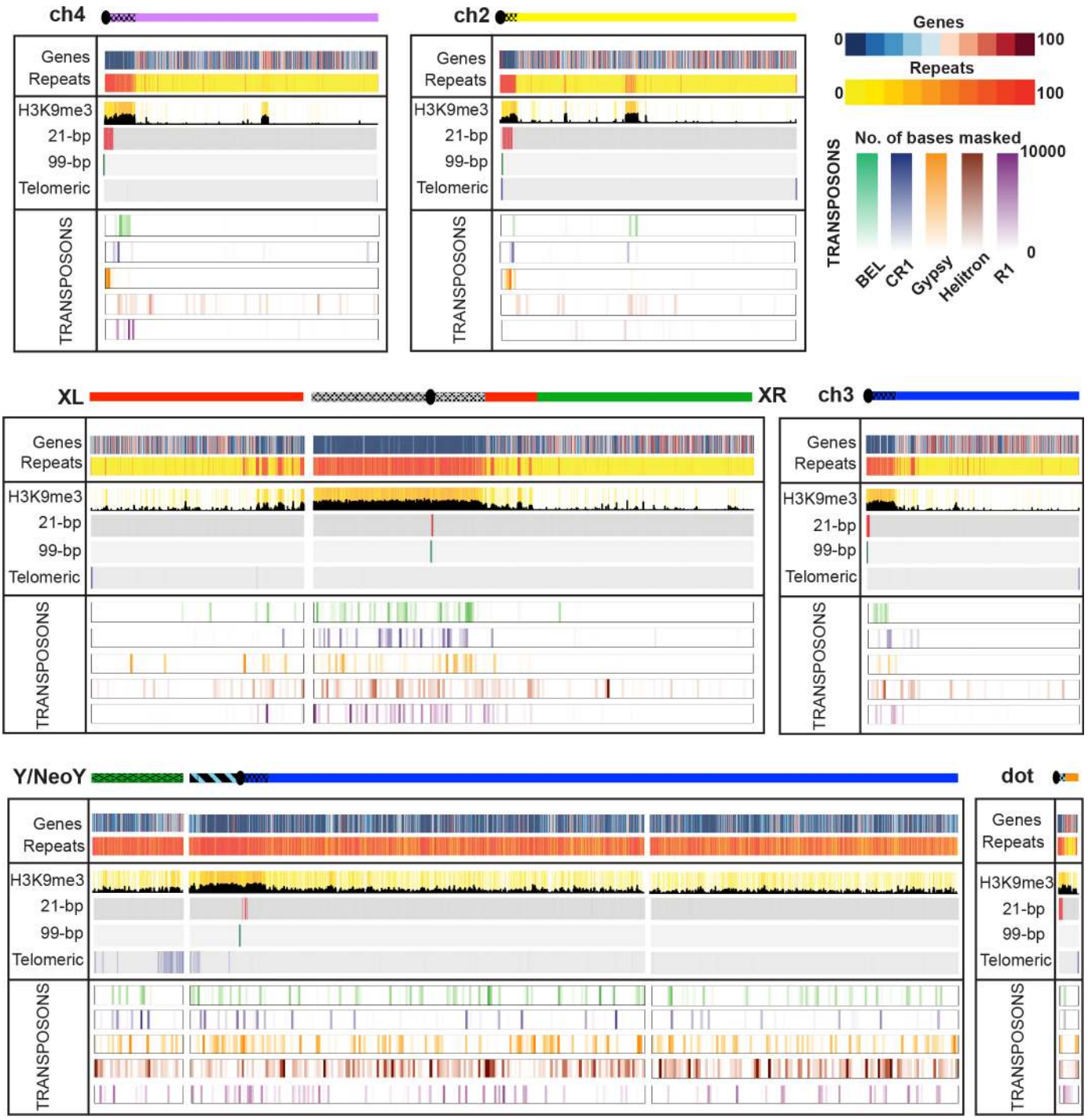

Fig 3. Gene and repeat content of *Drosophila miranda* genome assembly. Shown is the gene content, repeat content, H3K9me3 enrichment and density of the most abundant satellites (21-bp repeat and 99-bp repeat), the telomeric transposable elements, and the most abundant transposons (*BEL*, *CR1*, *Gypsy*, Helitron, *R1*) across the *D. miranda* genome assembly. A cartoon of the chromosomes is drawn, with color indicating the Muller element (see Fig 1D), and the shaded regions are heterochromatic. Gene and repeat content are shown in 40-kb sliding windows, and H3K9me3 enrichment and satellite and TE abundance are shown in 10-kb sliding windows. ch, chromosome; H3K9me3, trimethylation of histone 3 lysine 9.

https://doi.org/10.1371/journal.pbio.2006348.g003

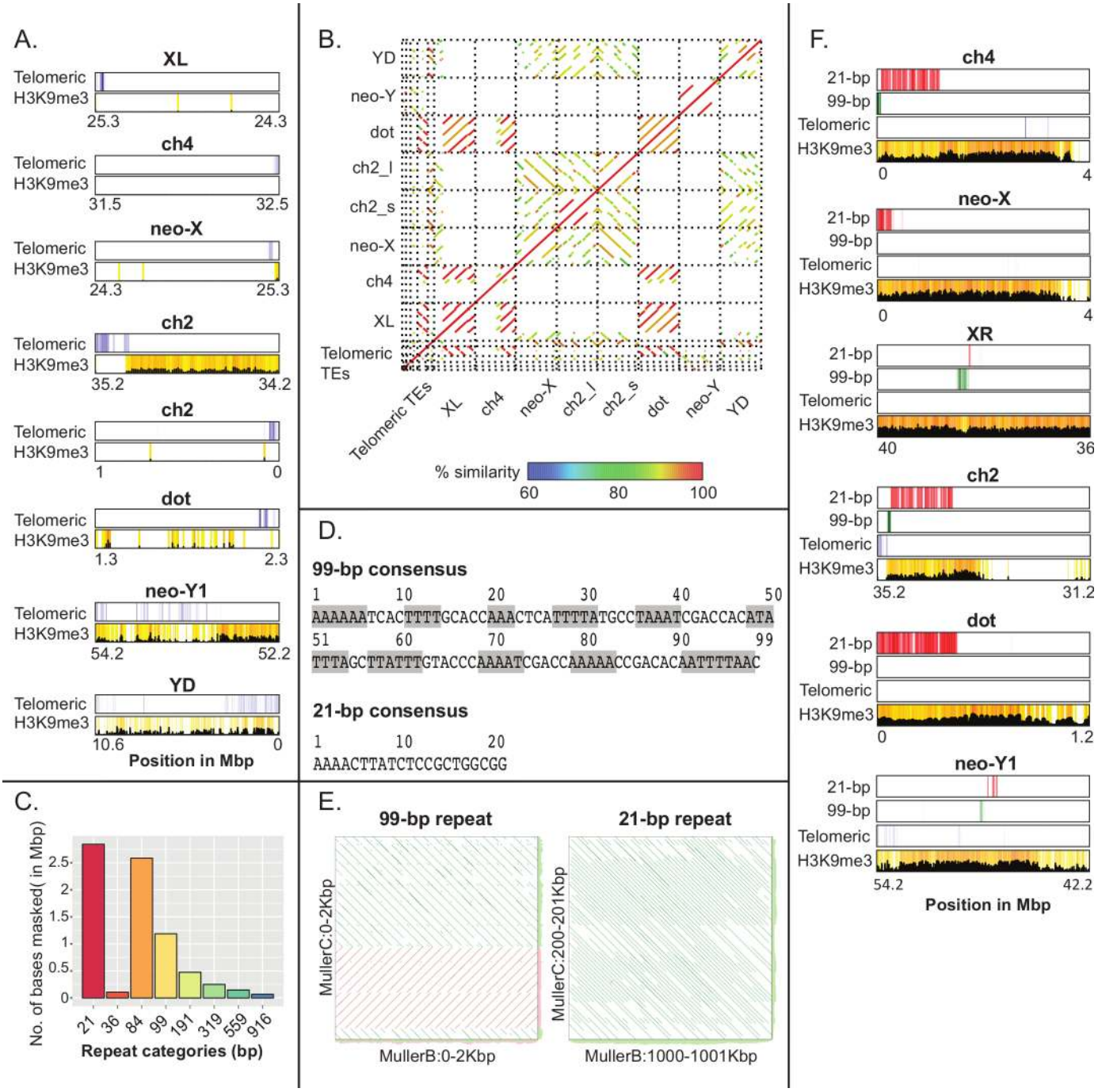

Fig 4. Recovery of telomeres and identification of putative centromere repeats for each chromosome. A. Presence of telomere repeats at or near the ends of most chromosome arms. Shown is enrichment of telomere repeats and H3K9me3 marks in 10-kb nonoverlapping sliding windows. **B.** Alignment of chromosome ends and telomere repeats. Colors indicate the percent similarity between the alignments and the direction of the lines indicates the direction of the match. **C.** Histogram of most abundant satellites in *Drosophila miranda* genome. Repeat categories refer to the size of the repeat unit. Note that the 84-bp repeat is a higher-order variant of four units of the 21-bp repeat. **D.** Consensus sequence of 21-bp and 99-bp repeats. Gray shading indicates AA/TT/AT repeats that occur at a 10-bp periodicity. **E.** Comparison of the centromeric repeat from different chromosomes. Shown are alignments of regions from Muller B and Muller C, with high density of the 99-bp and 21-bp tandem centromeric repeats, respectively. **F.** Location of putative centromere repeats in pericentromeric regions, and H3K9me3 enrichment. H3K9me3 enrichment is reduced at the putative centromeric repeats (S13 Fig). Note that, for the acrocentric chromosome 2, we recover the entire centromere, including the telomere. ch, chromosome; H3K9me3, trimethylation of histone 3 lysine 9; TE, transposable element.

https://doi.org/10.1371/journal.pbio.2006348.g004

comprise megabase-scale arrays of tandem repeats embedded in heterochromatin but are notoriously difficult to recover in genome assemblies. In several instances, we sequenced several megabases into the highly repetitive pericentromeric region (Fig 3, Fig 4; S11 Fig), and for one chromosome (Muller element E), we assembled the entire chromosome (based on the recovery of telomeric sequences on both chromosome ends), including its centromere.

We used Tandem Repeat Finder (TRF) [38] to identify satellite repeats, and plotted their occurrence along the genome (Fig 3, Fig 4C, S12 Fig). Interestingly, we find that the two most highly abundant repeats in the genome are adjacent to each other and heavily enriched along pericentromeric regions (Fig 4F, S13 Fig): a 21-bp motif that is found at the center of the centromeric region at most chromosomes, and an unrelated 99-bp repeat motif that is heavily AT-rich and has characteristics described for other centromeric repeats (Fig 4D, Fig 4E). Specifically, the 99-bp motif shows a 10-bp periodicity of A and/or T di- and trinucleotides, similar to centromere repeats found in diverse species, including *D. melanogaster* and the legume *Astragalus sinicus* [39,40]. Pericentromeric regions are heterochromatic, and we see strong enrichment of H3K9me3 along the pericentromere (Fig 3, Fig 4F). However, centromere repeats are partially occupied by a special centromeric variant of histone H3 (cenH3), which forms specialized nucleosomes that wrap centromeric DNA [41], and we would thus expect less H3K9me3 enrichment at sequences that partly replace the canonical H3 histone with cenH3. Indeed, H3K9me3 enrichment is reduced at both the 21-bp and 99-bp motif relative to other pericentromeric regions (Fig 4F, S13 Fig). Thus, the genomic distribution of the 21-bp and 99-bp motifs and their structural features and epigenetic modifications strongly suggest that they represent the functional centromere in *D. miranda*.

Repeat islands along euchromatic chromosome arms

In addition to the repetitive pericentromeres, our assembly also contains two large heterochromatic islands along the two autosomal arms (about 800 kb on chromosome [ch]4 and 1.5 Mb on ch2; Fig 3). These heterochromatic islands and their positions are supported by in situ hybridization data (Fig 1C). Intriguingly, while the repeat density is increased in these islands (especially on ch2), gene density is similar to other euchromatic regions. In *D. melanogaster*, repeat-rich heterochromatic regions appear to be absent along the major chromosome arms, and it will be of great interest to understand the functional significance and phylogenetic distribution of these heterochromatic islands.

Assembly of the Y and neo-Y chromosome of *D. miranda*

The presence of its recently formed neo-sex chromosomes has established *D. miranda* as an important model system [13,15,20]. Yet, the assembly of both the neo-X and neo-Y proved particularly challenging to short-read technology, and our previous attempt to create a contiguous Y/neo-Y chromosome assembly failed [20]. In contrast, our current assembly contains most of the Y chromosome in one large scaffold (101.5 Mb, see Fig 3, Fig 5). Intriguingly, the neo-Y assembly is about three times the size of the neo-X assembly (S7 Table, Fig 5); thus, analysis of neo-Y sequences based on neo-X alignments clearly misses the majority of the changes that occurred between the neo-sex chromosomes.

Sequence analysis of BAC clones confirms that our neo-X and neo-Y assembly is of high quality. In particular, 28 BAC clones fully map to the neo-X chromosome and 92 map to the neo-Y/Y chromosome; only three BACs in highly repetitive sequences on the neo-Y map to two different regions (and may either indicate a misassembly or a recombinant BAC clone; S5 Table, S6 Table; Fig 2D). Thus, our assembly approach allowed us to recover a highly contiguous Y/neo-Y sequence. Inspection of BAC sequences from homologous neo-X and neo-Y

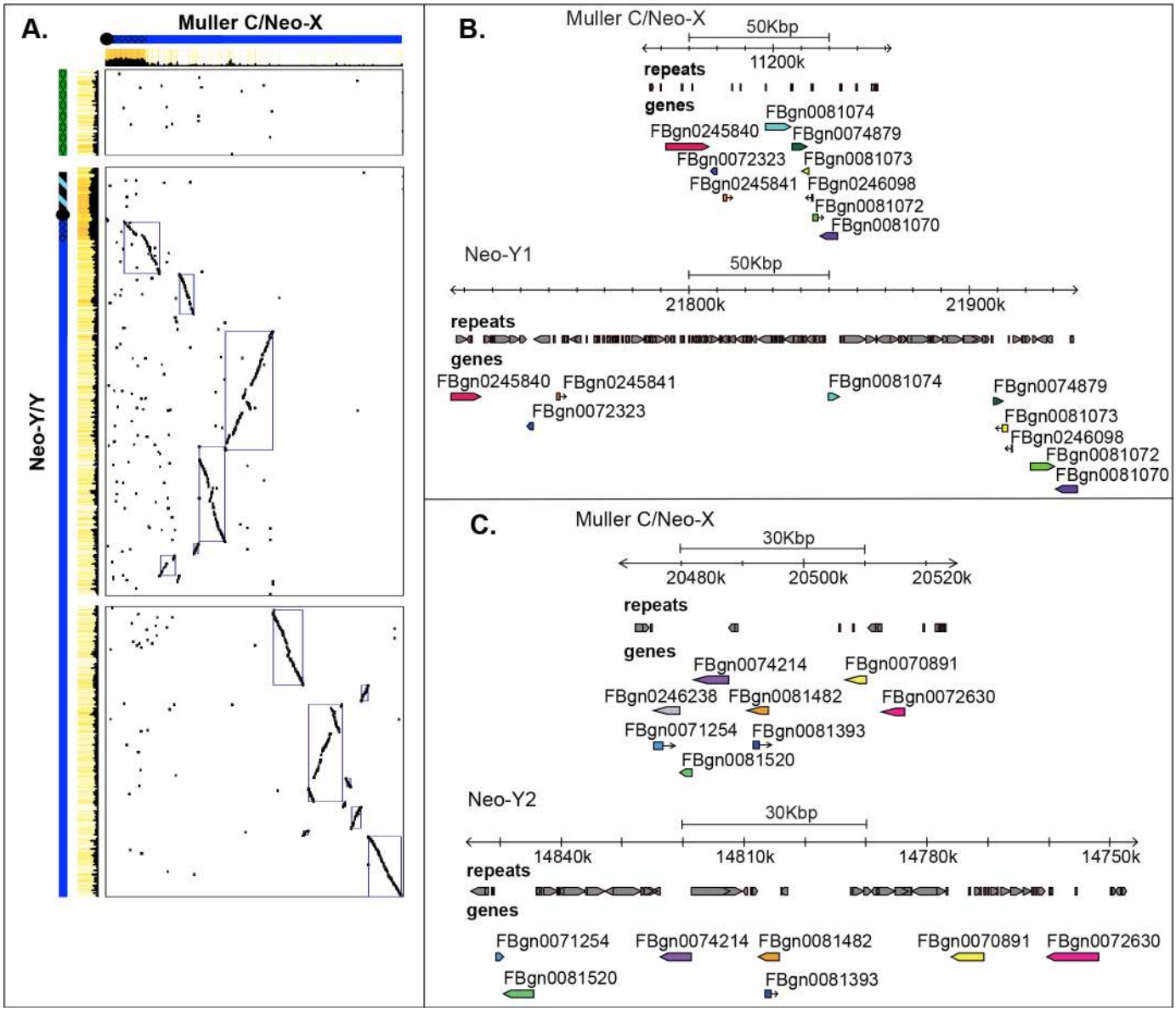

Fig 5. Neo-sex chromosome homology. A. Global neo-sex chromosome alignments show large homologous blocks between the neo-sex chromosomes along the long arm of the Y/neo-Y. **B** and **C**. Zoom-in of selected homologous regions along the neo-sex chromosomes. Neo-sex-linked regions often contain blocks of homologous genes, reflecting their recent evolutionary origin, but note the dramatic repeat accumulation (shown in gray) at both intergenic and gene regions on the neo-Y, greatly increasing its size.

https://doi.org/10.1371/journal.pbio.2006348.g005

regions confirms the specificity of our neo-X and neo-Y-linked assembly. That is, we find little cross-mapping between BAC clone sequences derived from the neo-X chromosome and its former homolog, the neo-Y, and vice versa, confirming the lack of chimeric assemblies (S14 Fig). Also, comparisons of homologous regions covered by BAC clones validate that neo-Y sequences contain roughly three times more DNA than their homologous segments on the neo-X, supporting the global size differences in chromosome assemblies that we observe. Thus, rather than shrinking—the fate that is typically ascribed to Y chromosomes—we find that early Y chromosome evolution instead is characterized by a massive global DNA gain.

Genome-wide alignments between the neo-X and neo-Y chromosome support a global increase in size of the neo-Y chromosome at intronic and intergenic regions, mainly driven by the accumulation of repetitive elements (Fig 3, Fig 5, S15 Fig). We assembled 110.5 Mb of Y/neo-Y-linked sequence, and 81.5 Mb are derived from repetitive elements (compared to 5.3 Mb on the neo-X). TEs are uniformly enriched along the neo-Y chromosome (Fig 3, Fig 5) and a main contributor to its dramatically increased genome size. Transposons show a highly nested structure on the neo-Y, with TE copies being disrupted by the insertion of (fragments of) other transposable elements (Fig 5B, Fig 5C), making the exact delineation of TE copies challenging. The most abundant repeat on the neo-Y is the *ISY* element [42], a helitron transposon that is inserted about 22,000 times on the neo-Y/Y chromosome and occupies more than 16 Mb on the neo-Y (i.e., 16% of neo-Y sequence; Fig 3, S16 Fig). In contrast, we only find about 1,500 copies on its former homolog, the neo-X chromosome (3% of the neo-X; less than 1 Mb). The second most common repeat class that has amplified on the neo-Y are *gypsy* transposable elements; we find roughly 14,300 insertions on the neo-Y (15% of neo-Y sequence, 15 Mb) and less than 1 Mb on the neo-X (about 800 insertions, i.e., 3% of its sequence).

Reconstruction of chromosomal events leading to sex chromosome turnover

In *D. miranda*, novel sex chromosomes were created recently at two different time points by chromosomal fusions (Fig 1D). In an ancestor of *D. miranda*, about 15 MY ago, a new X-linked arm (referred to as chromosome XR) arose by the fusion of an autosome (Muller element D) to the ancestral X chromosome (element A, referred to as chromosome XL in *D. miranda*). The fusion of Muller elements A and D left behind an unfused element D in males (which we refer to as YD), and this chromosome co-segregates with the ancestral Y and is transmitted through males only. The lack of recombination in male *Drosophila* implies that YD is entirely sheltered from recombination and thus undergoes genome-wide degeneration [12,43]. Indeed, while the fused Muller D that became part of the X chromosome has maintained most of its ancestral genes (we annotate over 2,800 genes on XR), previous attempts to recover Y-linked genes in *D. pseudoobscura* have proven difficult. On one hand, single-copy genes located on the ancestral Y chromosome (Y_{anc}) of *Drosophila* were found to be autosomal in *D. pseudoobscura* and its relatives [16], and located on the small dot chromosome (i.e., element F [17]). It was suggested that the current Y chromosome in *D. pseudoobscura* instead is the remnant of a highly degenerate YD [16], and we previously identified about 30 transcripts from the Y in *D. pseudoobscura*, most of which were found to have their closest paralogs on Muller D [44]. This supports the idea that the current Y of *D. pseudoobscura* is derived from the unfused Muller D, i.e., YD. A more recent chromosomal fusion (about 1.5 MY ago) between YD and element C created another neo-sex chromosome specific to *D. miranda*. This time, the fused element (termed the neo-Y) is male limited and undergoing degeneration, while the unfused element (the neo-X) is evolving characteristics typical of X chromosomes [19,20].

Our high-quality genome assembly allows us to reconstruct the evolutionary events leading to the independent creation of novel sex chromosomes in the *D. miranda* genome and the reversal of a former Y chromosome (Y_{anc}) to an autosome (Fig 6). Gene content is conserved across chromosomes in *Drosophila* (referred to as Muller elements A–F [45]), with Muller element A being the ancestral X chromosome in the genus *Drosophila* (Fig 1). We used orthology information from either *D. melanogaster* or *D. pseudoobscura* to infer chromosomal rearrangements in *D. miranda* and their evolutionary trajectory (Fig 6).

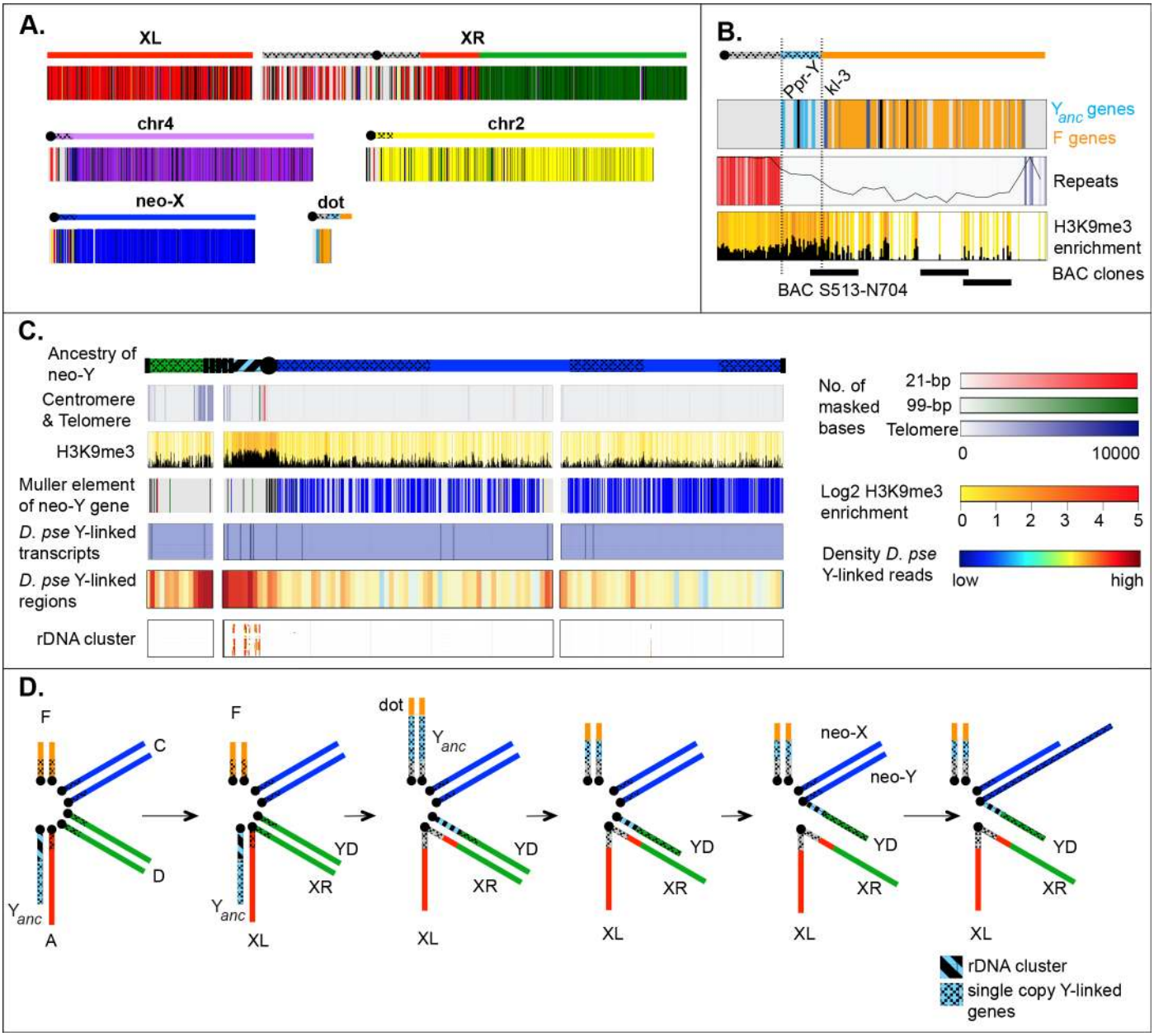

Fig 6. Karyotype evolution in *Drosophila miranda*. A. Chromosome arm homology in *D. miranda*. Genes in *D. miranda* are color coded according to their location in *D. melanogaster* (see Fig 1). **B**. Sequence composition of the *D. miranda* dot chromosome. Shown is the origin of dot genes (color coded as in Fig 1), the repeat and H3K9me3 content, as well as the location of sequenced BAC clones. *Ppr-Y* and *kl-3* are genes located on the ancestral Y of *Drosophila*. **C.** Origin of the *D. miranda* Y/neo-Y. Shown are the location of centromeric and telomeric repeats, H3K9me3 enrichments, the color coded location of single-copy neo-Y genes (with black corresponding to unknown ancestral location in *D. pseudoobscura*), the location of homologous Y-linked genes identified in *D. pseudoobscura*, mapping of Y-derived sequencing reads from *D. pseudoobscura*, and the location of the rDNA genes. The inferred ancestry of the Y/neo-Y chromosome is shown as a cartoon, with the short arm presumably corresponding to the Y chromosome shared with *D. pseudoobscura* and the long arm representing the neo-Y. **D.** Our genomic analysis allows us to reconstruct sex chromosome evolution in *D. miranda* (see text). BAC, bacterial artificial chromosome; chr, chromosome; H3K9me3, trimethylation of histone 3 lysine 9; rDNA, ribosomal DNA; Y_{anc}, ancestral Y chromosome.

https://doi.org/10.1371/journal.pbio.2006348.g006

The fusion between elements A and D created a metacentric X chromosome in *D. miranda*, and our assembly contains both of these chromosome arms as a single scaffold, including a large pericentromeric block on XR that is highly repeat rich. Comparison to *D. melanogaster* identifies a pericentric inversion that moved approximately 340 genes from element A onto

XR (see Fig 6A). An X chromosome–autosome fusion results in two Y chromosomes in males (i.e., Y_{anc} and the unfused element D), but *D. miranda* and its relatives only harbor a single Y. The ancestral Y chromosome in *Drosophila* contains a handful of single-copy genes that have no homologs on the X chromosome [46,47], as well as the multi-copy ribosomal DNA (rDNA) repeat cluster that is present on both the X and the Y and used for pairing of the sex chromosomes during male meiosis [48–50]. Our assembly reveals that the gene content of the ancestral Y is split up between two chromosomes: all five ancestral single-copy Y genes in *Drosophila* (i.e., *kl-2*, *kl-3*, *ORY*, *PRY*, and *PPr-Y*) are located in a single genomic region on the dot chromosome, adjacent to the centromere (Fig 6B), while the rDNA repeat cluster is found on the Y chromosome of *D. miranda* (Fig 6C). The presence of two Y chromosomes in an ancestor of *D. miranda* may have resulted in an increased frequency of aneuploidy gametes [51]. Potential problems in meiosis could have been ameliorated by the fusion and/or translocation of genetic material from Y_{anc} to both element F and YD, and relocation of the rDNA repeat cluster onto YD could have helped to ensure proper segregation between the X and Y chromosome. Indeed, an in situ hybridization study suggests that copies of the rDNA loci exist on both the X and Y chromosomes in relatives of *D. miranda* that share the element A–D fusion and translocation of single-copy Y_{anc} genes onto element F [17], suggesting that these structural rearrangements co-occurred rapidly before the divergence of this species group. Note, however, that the chromosomal location of the rDNA cluster can differ among closely related *Drosophila* species [52], so other scenarios of movement of rDNA genes are possible.

The Y-derived material on the dot of *D. miranda* amounts to approximately 300 kb, which is substantially smaller than Y chromosomes found in *Drosophila* [53], suggesting that Y_{anc} presumably lost genetic material after fusing to the dot chromosome. Similar shrinkage of the Y_{anc} was found in its sister species, *D. pseudoobscura* [18], which shows an inversion of the Y-derived segment with respect to *D. miranda* (S17 Fig). The Y_{anc}/element F fusion break point is corroborated independently by a BAC clone spanning the fusion (S17 Fig), validating our genome assembly in this region. Despite Y_{anc} presumably having lost large amounts of repetitive DNA, we find its repeat content to be elevated relative to euchromatic regions, and Y_{anc} genes contain higher levels of heterochromatin compared to genes from other chromosomes (S17 Fig). They also have maintained their testis-specific expression pattern in *D. miranda* (S17 Fig). Thus, despite having become linked to an autosome, single-copy Y genes have retained their ancestral chromatin environment and testis function.

Nonrecombining Y chromosomes degenerate within a few MY in *Drosophila* [19,20], and most ancestral genes on YD were presumably lost before it fused to element C about 1.5 MY ago. We tried to reconstruct the evolutionary history of the Y chromosome in *D. miranda* by identifying which parts of the Y/neo-Y chromosome were derived from Muller D versus Muller C versus the original Y_{anc}. Our Y/neo-Y chromosome assembly consists of two chromosome arms, spanning the putative centromeric repeats, and the heterochromatic pericentromere (Fig 6C). A dot plot between the neo-X (Muller C) and neo-Y reveals several large blocks of homology on the large Y/neo-Y arm but none on the shorter arm (Fig 5A). Fig 6C plots the location of single-copy genes along the neo-Y/Y chromosome of *D. miranda*, color coded by Muller element. We identify many genes from the long arm of the Y/neo-Y, most of which are derived from Muller C; in contrast, only few unique genes exist on the short arm, and their closest homologs are not preferentially located on Muller C (Fig 6C). This suggests that the long arm is derived from the neo-Y, but not the shorter one, which instead may be derived from YD and should thus also be Y-linked in *D. pseudoobscura*. The current genome of *D. pseudoobscura* lacks an assembly of its Y chromosome, and repetitive nonfunctional regions evolve rapidly, which makes identification of YD sequences challenging. We attempted to detect putative YD sequences by identifying reads and scaffolds from the fragmented *D.*

pseudoobscura genome that are male specific (see Methods) and mapping them onto our *D. miranda* Y/neo-Y assembly. Preferential mapping of putative male-specific (i.e., Y-linked) sequences from *D. pseudoobscura* to the short arm of the *D. miranda* Y/neo-Y chromosome assembly supports the notion that the short arm of the Y/neo-Y chromosome corresponds to YD. The rDNA cluster maps adjacent to the centromere on the short arm of the Y chromosome, which suggests that this part is derived from the original Y (i.e., Y_{anc}) of *Drosophila*.

Interestingly, hybridization studies have shown that the Y/neo-Y chromosome of *D. miranda* contains about 70 copies of the telomere repeat [34] and displays an intensely labeled internal telomere-repeat block adjacent to the centromere [54]. Indeed, our assembly recovers a large internal block of telomere-repeat sequences close to the centromere (**Fig 6C**), bordering fragments of the Y chromosome of different evolutionary origin (i.e., they are found between fragments derived from Muller D versus Muller C versus the original Y_{anc}). Telomere repeats within the Y/neo-Y may present the remnants of a "telomere-to-telomere" type chromosomal fusion that created the neo-Y/Y chromosomal arrangements in this species.

Conclusion

Here, we create a genome assembly of unprecedented quality and contiguity for the fruit fly *D. miranda*, a species that has served as a model for sex chromosome research. In *D. miranda*, chromosomal fusions at different time points independently created de novo sex chromosomes or led to the reversal of a former Y to an autosome, and our high-quality assembly allows us to reconstruct the evolutionary events creating and dismantling sex chromosomes. Our assembly recovers entire chromosomes and notoriously difficult regions to assemble, including entire centromeres or large stretches of repetitive sequences, such as the rDNA cluster. All chromosome arms of *D. miranda*—including its Y chromosome—are contained in a single, chromosome-sized scaffold, and in almost all cases, chromosome arms are flanked by telomere sequences on one end and centromeric repeats on the other. This demonstrates that long molecule sequencing approaches have great potential to assemble highly repeat-rich regions, such as Y chromosomes and centromeres [55–57], which will allow studying the function, biology, and evolution of repetitive regions in many species, including gene family expansions and contractions, identification and characterization of centromeres, heterochromatin function, genomic analysis of Y chromosomes, repeat evolution, or identification of novel genes embedded in heterochromatin. In one instance, we assemble an entire chromosome, fully sequence through the pericentromeric DNA and the centromere, and recover telomeres on both ends. Our high-quality assembly allows us to infer the centromeric satellite DNA motif in *D. miranda*, which shares no sequence similarity with other centromeres but has characteristics typical of centromeric repeats, including a 10-bp periodicity of AA/TT/AT repeats. This sequence feature presumably helps to stabilize centromeric nucleosomes that may be under tension during anaphase, because a single turn of the DNA double helix is approximately 10 bp, and sequences with 10-bp periodicity in AA, TT, or AT dinucleotides favor wrapping of nucleosomes by reducing the bending energy of wrapping [58,59]. Lack of sequence conservation of centromeric repeats confirms that centromeres turn over quickly [37], and will allow the functional characterization and investigation of centromere biology in this group. For the first time, we also assemble an entire Y chromosome using shotgun sequencing approaches. In particular, the recovered Y/neo-Y sequence is over 100 Mb large, which is over three times the size of that of its former homolog, the neo-X, or its autosomal ortholog in *D. pseudoobscura*. Thus, rather than shrinking—the fate that is typically ascribed to animal Y chromosomes—we find that early Y chromosome evolution instead is characterized by a global DNA gain. Large young Y chromosomes have been observed in plants, and like in

D. miranda, their length increase is primarily due to an accumulation of repetitive DNA [60,61]. We show that the *D. miranda* Y chromosome provides a hodgepodge of sequences that have been male limited for different amounts of time, and display various stages of degeneration. The ancestral Y chromosome of *Drosophila*, on the other hand, has become linked to an autosome in *D. miranda*, and we reconstruct the genomic and epigenetic changes that occurred to revert this former Y to an autosome. Thus, our new highly improved genome assembly will provide the basis for further evolutionary and functional research on repetitive sequences and the recently formed neo-sex chromosomes of *D. miranda*.

Methods

Fly strain

We chose the inbred MSH22 strain for *D. miranda*, which was previously used to generate a BAC library [19], and for genome assembly using short Illumina reads [20].

PacBio DNA extraction and genome sequencing

We used a mix of MSH22 males and extracted high molecular weight DNA using a QIAGEN Gentra Puregene Tissue Kit (Cat #158667), which produced fragments >100 kbp (measured using pulsed-field gel electrophoresis). DNA was sequenced on the PacBio RS II platform. In total, this produced 28 Gb spanning 2,407,465 filtered subreads with a mean read length of 12,818 bp and an N50 of 17,116 bp (**S1 Table**, **S18 Fig**).

BioNano DNA extraction and optical mapping

DNA was extracted from flash frozen male larvae. Purified DNA was embedded in a thin agarose layer and was labeled and counterstained following the IrysPrep Reagent Kit protocol (BioNano Genomics). Samples were then loaded into IrysChips and run on the Irys imaging instrument (BioNano Genomics). This produced 90,977 molecules (molecule length: minumum 150,000, median 191,400, and maximum 1,957,000 and N50 of 209,014; **S9 Table**; **S19 Fig**). The IrysView (BioNano Genomics) software package was used to produce single-molecule maps and de novo assemble maps into a genome map (**S4 Table**). The BioNano assembly has 401 contigs with an N50 of 0.5 Mb and assembled length of about 178 Mb. HybridScaffold was then used to produce hybrid maps from the BioNano contigs and the genomic scaffolds from our scaffolded PacBio assembly, and IrysView was used to visualize alignments of the BioNano contigs and genomic scaffolds to the hybrid ones. **S4 Table** shows coverage of hybrid scaffolds by BioNano contigs and NGS contigs (genomic scaffolds).

PacBio assembly

An initial PacBio assembly was built with the Falcon assembler [23], using 40× error corrected reads. Twenty-eight-Gb of long reads (NR50 = 17,116 bp; NR50 is the read length, such that 50% of the total sequence is contained within reads of this length or longer) were assembled using Falcon assembler (v1.7.5) [23] running on Sun Grid Engine in parallel mode. For assembly, reads longer than 10 kb and 17 kb were used as seed reads for initial mapping and preassembly. The options for read correction, overlap filtering, and consensus building were provided in the config file as follows: pa_HPCdaligner_option = -v -dal128 -t16 -e.70 -l1000 -s1000; ovlp_HPCdaligner_option = -v -dal128 -t32 -h60 -e.96 -l500 -s1000; pa_DBsplit_option = -x500 -s400; ovlp_DBsplit_option = -x500 -s400; falcon_sense_option = —output_multi—min_idt 0.70—min_cov 4—max_n_read 200—n_core 6; overlap_filtering_setting = —max_diff 30—max_cov 60—min_cov 5—n_core 24. This assembly had 629 scaffolds and a total assembled length of

274,803,116 bp with an N50 value equal to 2,188,952 bp. We polished this assembly using the software Quiver [62], followed by the software Pilon [63], which resulted in an assembly with 625 scaffolds, with an N50 value of 2,232,625 bp and total assembled length equal to 271,223,447 bp. We also produced a second PacBio assembly using Canu [24], with default parameters. This assembly consisted of 521 scaffolds and a total assembled length of 296,012,170 bp, with an N50 value of 3,884,273 bp. The Canu and the Falcon assemblies both contained some regions that were missing from the other one, and the two assemblies were merged using Quickmerge [25], with default parameters. The resulting merged assembly was then merged a second time to the finished Falcon assembly, producing a superior Quickmerge assembly consisting of 271 scaffolds and total length equal to 295,213,648 bp and an N50 value of 5,177,776 bp.

Hi-C libraries

Hi-C libraries were created from sexed male and female third instar larvae of MSH22, following [64]. Briefly, chromatin was isolated from male and female third instar larvae of *D. miranda*, fixed using formaldehyde at a final concentration of 1%, and then digested overnight with HindIII and HpyCH4IV. The resulting sticky ends were then filled in and marked with biotin-14-dCTP, and dilute blunt end ligation was performed for 4 hours at room temperature. Cross-links were then reversed, and DNA was purified and sheared using a Covaris instrument LE220. Following size selection, biotinylated fragments were enriched using streptavidin beads, and the resulting fragments were subjected to standard library preparation following the Illumina TruSeq protocol. For females, 38.4 and 194.5 million 100-bp read pairs were produced for the HpyCH4IV and HindIII libraries, respectively. For males, 28.0 and 179.2 million pairs were produced.

Hi-C-based proximity-guided (PG) assembly

We mapped Illumina male and female genomic paired-end reads and classified contigs as autosomal, X-linked, or Y-linked based on genomic coverage. We created two pools of contigs: autosomes or X-linked, and Y-linked, and scaffolded them separately. We used Juicer [65] to align female Hi-C reads to the autosomal/X-linked scaffolds and also to align a subset of male Hi-C reads (that did not map to autosomes) to the Y-linked scaffolds. There were 22,168,695 Hi-C contacts: 2,921,250 interchromosomal and 19,247,445 intrachromosomal contacts for the autosomal/X-linked scaffolds. For the Y-linked scaffolds, there were 795,487 Hi-C contacts, including 173,147 interchromosomal and 622,340 intrachromosomal contacts. The output alignment files from Juicer were then used to scaffold the genome using 3D-DNA [66]. Using a custom Perl script, we then scaffolded the PacBio assembly fasta based on the 3D-DNA output suffixed .asm, which contains information about the positions and orientations of contigs; scaffolded contigs are gapped by 50 Ns. With the Hi-C scaffolded assembly, we then realigned the Hi-C reads using bwa mem [67] single-end mode on default settings. The resulting *sam* files were then used to generate a genome-wide Hi-C interaction matrix using the program Homer [68]. For visualization, we plotted the interaction matrix as a heatmap in R, with demarcations of the PacBio contigs and Hi-C scaffolds. Iteratively, we visually examine the heatmap to identify possible anomalies as scaffolding errors and manually curate the .asm file output to improve the heatmap. At each stage of the assembly process, genome completeness was assessed using BUSCO (v 3.01) [29], using the arthropod database (odb9).

BAC clone DNA isolation and sequencing

Bacteria were cultured in Terrific Broth with 25 μg/mL chloramphenicol. Overnight cultures (500 μL) were inoculated with starter cultures grown from glycerol stocks, covered with Area-Seal films, and incubated at 37 ˚C with shaking for 12–14 hours. Overnight cultures were

pelleted by centrifugation, resuspended in 60 μL [Tris-HCl (50 mM, pH 8) and EDTA (50 mM)], and lysed by adding 120 μL [NaOH (200 mM) and SLS (1%)]. Cells were incubated at room temperature for 5 minutes, 270 μL [KOAc (5 M, pH 5)] was added and chilled on ice for 10 minutes, and then centrifuged for 1 hour. DNA was precipitated with isopropanol, washed with 70% and 80% ethanol and eluted in Qiagen EB (50 μL). Nextera libraries were prepared from the BAC DNA, following Illumina's protocol with the following modifications: reaction volumes were scaled to 1 μL input BAC DNA (@ 1–3 ng/μL), and SPRI bead cleanup steps after tagmentation and PCR amplification were skipped. Barcoded libraries were pooled, and a two-sided Ampure XP size selection removed fragments <200 bp and minimized fragments >800 bp. The pooled libraries were sequenced on a HiSeq 4000 with 100-bp paired-end reads.

BAC clone mapping

For each BAC clone, Nextera reads were first adapter trimmed using cutadapt (http://code.google.com/p/cutadapt/) and filtered to remove concordantly mapping read pairs from pTARBAC-2.1 and *E. coli* DH10B using Bowtie2 [69]and SAMtools [70]. The remaining trimmed, filtered reads were mapped to our *D. miranda* assembly using bwa [67]. The BAC's location was determined by filtering regions of high coverage (at least 50× mean) and significant length (at least 20 kb). First, regions with average coverage of at least 50× were extracted, and any regions within 250 kb of each other were merged using BEDtools [71]. When this resulted in a merged region longer than 250 kb, the merging step was repeated on this long region using a maximum distance of 5 kb. If only one region remained, this was defined as the putative BAC location. If multiple regions were found, they were ranked by average coverage, and any region with less than half the average coverage of the region with the highest average coverage was considered cross contamination. Finally, regions less than 20-kb long were removed.

To confirm that reads mapping to these BAC locations included both edges of the BAC insert, we found discordantly mapping read pairs with one read mapping to the vector and its mate mapping to our assembly. Filtered reads were mapped to pTARBAC-2.1 using bwa [67], and discordantly mapping reads from either end were filtered from the .sam file, keeping "start" and "end" reads separated (reads mapping to a region within 4,000 bp of the vector's start position were considered "start" reads, and reads mapping within 4,000 bp of the vector's end position were considered "end" reads). The mates of these start/end reads were extracted, merged, and counted using BEDtools [71] and filtered to find edge read pileups within 10 kb of the putative BAC edges. To confirm that these edge reads are at either end of each BAC location, IGV snapshots with three tracks (all mapped reads, "start" reads, and "end" reads) were reviewed manually.

To confirm that our assembly of the neo-X and neo-Y were highly specific and accurate, the genomic region on the neo-sex chromosome from which a specific BAC clone was derived was masked using BEDtools [71], and the BAC clone reads were mapped back to this masked assembly and then filtered and merged, as described above. Regions of primary and secondary mapping were reviewed using IGV to confirm that little cross-mapping occurs in our assembly; after masking and remapping, we found significant mapping to homologous regions of its homologous neo-sex chromosome, but mapped reads typically contained many SNPs and many gapped regions (**S14 Fig**).

Conflict resolution

To identify large-scale, erroneously duplicated regions, we took advantage of the fact that when reads are mapped equally well to multiple regions, they are randomly assigned to one of the regions; we mapped Illumina reads to the assembly twice and identified >100-kb regions where roughly half of the reads map to another region in the two mappings (see **S4 Fig**). For

erroneous duplications and mis-scaffolded contigs in the PacBio assembly identified, we used IGV to visualize the quality of Illumina reads mapping, in order to determine the precise coordinates to modify our assembly (**S2 Table**). For erroneous duplications, we identified the position in which Illumina reads are no longer uniquely mapping around the duplicated areas; one of the two duplications is then removed. Mis-scaffolded contigs are typically caused by misassembly around repetitive elements; therefore, we also relied on visual inspection of nonuniquely mapping reads to separate contigs.

Structural variant calling for quality control

For the previously published genome assembly and the various intermediate assemblies produced here during generating the current version, we estimated quality statistics using the variant caller LUMPY [72]. To do this, we first aligned reads from two separate male Illumina libraries (with 626-bp and 915-bp insert sizes, respectively) to our current assembly and its intermediates using SpeedSeq, which does a BWA-MEM alignment and produces discordant and split reads bam files. We ran the software lumpyexpress [72] using these bam files, which produced a vcf file with several categories of structural variants: BND = trans-contig associations, DEL = deletions, DUP = Duplications, INV = Inversions. High numbers of these variants are indicative of potential assembly errors and provide a meaningful way to assess assembly quality.

Repeat annotation and masking

For repeat masking the genome, we annotated repeats using REPdenovo (downloaded November 7, 2016 [26]) and RepeatModeler version 1.0.5 [27]. We ran REPdenovo on raw sequencing reads using the parameters MINREPEATFREQ 3, RANGEASMFREQDEC 2, RANGEASMFREQGAP 0.8, KMIN 30, KMAX 50, KINC 10, KDFT 30, GENOMELENGTH 176000000, ASMNODE LENGTHOFFSET -1, MINCONTIGLENGTH 100, ISDUPLICATEREPEATS 0.85, COVDIFF CUTOFF 0.5, MINSUPPORTPAIRS 20, MINFULLYMAPRATIO 0.2, TRSIMILARITY 0.85, and RMCTNCUTOFF 0.9. We ran RepeatModeler with the default parameters.

We used tblastn (https://www.ncbi.nlm.nih.gov/BLAST/) with the parameters -evalue 1e-6, -numalignments 1, and -numdescriptions 1 to blast translated *D. pseudoobscura* genes (release 3.04) from FlyBase [73] to both (REPdenovo and RepeatModeler) repeat libraries. We eliminated any repeats with blast hits to *D. pseudoobscura* genes. After filtering, our REPdenovo repeat annotation had 999 repeats totaling 964,435 base pairs.

We also made a REPdenovo annotation using a subset of female reads, for which we also filtered out repeats blasting to *D. pseudoobscura* genes. This annotation had 716 repeats totaling 544,702 base pairs. We used RepeatMasker version 4.0.6 [27] and blastn (https://www.ncbi.nlm.nih.gov/BLAST/) with the parameters -evalue 1e-6, -numalignments 1, and -numdescriptions 1 to blast this annotation to the Repbase *Drosophila* repeat annotation (downloaded March 22, 2016, from www.girinst.org) in order to classify repeats from this annotation. Our RepeatModeler repeat annotation had 1,009 repeats totaling 1,290,513 base pairs. Of the 1,009 repeats, 103 were annotated as DNA transposons, 145 as LINEs, 365 as LTR transposons, 42 as Helitrons, and 1 as a SINE. We concatenated our filtered REPdenovo and RepeatModeler repeat annotations to repeat-mask the genome with RepeatMasker [74].

Gene annotation using Maker

To run Maker [28], we first build transcriptome assemblies. RNA-seq reads from several adult tissues (male and female heads, male and female gonads, male accessory gland, female spermatheca, male and female carcass, male and female whole body, and whole male and female third instar larvae; see **S10 Table**) were aligned to the genome assembly using HiSat2 [75], using

default parameters and the parameter -dta needed for downstream transcriptome assembly. The alignment produced by HiSat2 was then used to build a transcriptome assembly using the software StringTie [76] with default parameters, which produced a transcript file in gtf format. Fasta sequences of the transcripts were then extracted using gffread to be used with Maker. The genome was repeat-masked using RepeatMasker and our de novo repeat library as well as the Repbase (http://www.girinst.org/) annotation.

We ran three rounds of Maker [28] to iteratively annotate the genome. For the first Maker run, we used annotated protein sequences from FlyBase for *D. melanogaster* and *D. pseudoobscura* as well as the de novo assembled *D. miranda* transcripts and the genes predictors SNAP [77] and Augustus [78] to guide the annotation. We used the SNAP *D. melanogaster* hmm and the Augustus fly model, with the parameters est2genome and protein2genome set to 1 in order to allow Maker to create gene models from the protein and transcript alignments. Before running Maker a second time, we first trained SNAP using the results of the previous Maker run and set the est2genome and protein2genome parameters to 0. We then used our new hmm file and the Augustus fly model to annotate the genome. The third iteration was done similarly to the second one by training SNAP on the results of the previous Maker run. This procedure resulted in a total of 17,745 annotated genes. The repeat and gene densities were plotted for the major chromosomal arms and scaffolds using the software DensityMap [79].

Tandem repeat identification and quantification

We used TRF [38] on recommended settings to identify tandemly repeating motifs across the assembly. To identify variants or multimers of the same motif, the identified motifs are then blasted pairwise to themselves. Those that are over 90% identical for over 90% of the length are grouped together and collapsed into the same motif. Satellites' abundances were parsed from the TRF output and RepeatMasker output using the identified motifs as the repeat library.

Identifying telomeric repeats

Telomeric protein sequences for *D. pseudoobscura* and *D. persimilis* from [33] were aligned to the de novo repeat library using BLAST. Hits with a score greater than 50 and percent identity greater than 75 were classified as telomeric and RepeatMasker was used to identify their genomic locations. A heatmap showing the number of bases masked in 10-kb windows was then plotted along the genome using R.

Identifying orthologous proteins and whole genome alignments

We identified orthologous proteins by aligning *D. pseudoobscura* proteins to our list of de novo annotated *D. miranda* proteins using BLAST and BLAT. For 16,378 of the total 17,745 genes in our annotation, we were able to reliably identify orthologs in the *D. pseudoobscura* annotation. We used blastp to align protein sequences of the remaining 1,367 genes to annotated *D. melanogaster* proteins and were able to identify *D. melanogaster* orthologs for 285 of these 1,367 genes. Thus, we were unable to identify orthologs for 1,082 genes in both the *D. pseudoobscura* and the *D. melanogaster* genome. Whole genome alignments were performed using Nucmer (from the MUMmer package [80]) and dot plots were produced using mummerplot, symap42 [81], or YASS [82].

Identifying *D. pseudoobscura* Y-linked reads

Scaffolds from a male-only *D. pseudoobscura* assembly were aligned to a female-only *D. pseudoobscura* assembly using Nucmer (from the MUMmer package [80]) to identify scaffolds

only present in the male assembly (i.e., putative Y-linked scaffolds). Male Illumina reads were then aligned to these scaffolds using bowtie2 [69] and unaligned reads were discarded. The aligned reads were then mapped to the female genome and any reads that mapped were discarded to further enrich for only male-specific reads. These reads were then mapped to the *D. miranda* Y/neo-Y-linked scaffolds, and coverage was calculated in 10-kb nonoverlapping windows. The density of nonzero coverage windows was plotted along the three largest Y scaffolds.

Supporting information

S1 Fig. Illumina sequencing coverage of three individual females (red) and males (blue), and the female-to-male coverage ratio (black). The chromosomal scaffolds (after Hi-C scaffolding) and unscaffolded contigs are demarcated by dotted lines and ordered based on their female-to-male coverage ratio. Each dot represents the average coverage across a 50-kb window.
(PDF)

S2 Fig. Hi-C association density maps for autosomes and X contigs, and Y contigs. The Hi-C association heatmap of the PacBio contigs (demarcated by dotted lines), sorted by contig size (left panels), is reorganized using 3D-DNA (middle panels), generating near-chromosome-length scaffolds (black boxes, right panels).
(PDF)

S3 Fig. Hi-C linkage density map, gene and repeat content for **A.** chromosome XL, **B.** chromosome XR, **C.** chromosome 2, **D.** chromosome 4, **E.** neo-X chromosome, **F.** chromosome YD, **G.** neo-Y_1, **H.** neo-Y_2, **I.** Muller F. Neo-Y_1 and neo-Y_2 refer to the two largest neo-Y scaffolds (see Fig 2C). Note that regions of increased repeat density (such as centromeres or the repeat islands on chromosome 2 and 4) show increased contact probabilities with other repeats. Chromosome arms/scaffolds are not drawn to scale.
(PDF)

S4 Fig. Erroneous duplications in the PacBio assembly. The normalized female and male Illumina sequence read coverages along the PacBio assembly are plotted in the outer circles in red and blue, respectively. Duplications greater than 100 kb in the assembly are connected with black lines. Erroneously duplicated regions are accompanied by sharp reduction of the coverage by half, because of the sequencing reads being divided between the erroneously duplicated regions (examples marked by arrows). True duplications will show no reduction in read depth.
(PDF)

S5 Fig. Validation of our assembly using BioNano optical maps. Shown are alignments of BioNano contigs and NGS scaffolds (PacBio and Hi-C scaffolds) to hybrid scaffolds, and alignments of BioNano molecules to HybridScaffolds for the different chromosomes. Chromosome arms/scaffolds are not drawn to scale. NGS, next-generation sequencing.
(PDF)

S6 Fig. Mapping of some BAC clones. We evaluated IGV plots for all sequenced BAC clones to confirm that they map contiguously and uniquely, and we identified reads mapping to the edge of BAC clones (indicated by red reads). BAC, bacterial artificial chromosome; IGV, integrative genomics viewer.
(PDF)

S7 Fig. Comparison of current (Dmir2.0) versus old (Dmir1.0) *D. miranda* assembly for **A.** chromosome XL, **B.** chromosome XR, **C.** chromosome 2, **D.** chromosome 4, and **E.** neo-X chromosome. Note that Dmir1.0 contains dozens of inversions that were probably introduced by scaffolding contigs with the *D. pseudoobscura* genome assembly. Also, Dmir1.0 is substantially shorter, mainly because of the almost complete absence of repetitive sequences from this assembly, such as pericentromeres.
(PDF)

S8 Fig. Comparison of current *D. miranda* assembly (Dmir2.0) versus *D. pseudoobscura* assembly.
(PDF)

S9 Fig. Satellite DNA in the *D. melanogaster* assembly (r6). A. Distribution of centromeric satellites and telomeric retrotransposons (color labeled) are plotted along the scaffolds. **B.** The first (left) and last (right) 1 Mb are plotted for each chromosome arm. Note that while the *D. melanogaster* chromosome arms typically have their telomeres assembled (red repeats), centromeric repeats (all other repeats) are generally missing from the assembly.
(PDF)

S10 Fig. Comparative alignments of resolved tandemly duplicated gene clusters in *D. miranda*. Many tandemly duplicated regions were poorly represented in the published assembly and generally collapsed into a single copy.
(PDF)

S11 Fig. BAC clone sequencing confirms centromere assembly. BAC clone S506-N718 on Muller E is located in the middle of a 21-bp repeat region, supporting that our assembly is of high quality in the repeat-rich centromere and pericentromeric regions. BAC, bacterial artificial chromosome.
(PDF)

S12 Fig. Satellite DNA in the assembly. A. Distribution of satellites and telomeric retrotransposons (color labeled) are plotted along the scaffolds. The names of satellites are derived from the length of the base motif. Note that the 84-bp is a complex structure of four 21-bp variants. **B.** The first and last 1.5-Mb is plotted for select chromosomes. **C.** The abundances of each satellite type across the assembly. **D.** The number of base pairs masked by each satellite type. Underlying data can be found in S1 Data.
(PDF)

S13 Fig. Enrichment of H3K9me3 at pericentromeric regions and putative centromeric repeat for different chromosome arms (note that X–L and YD show no large regions containing the 21-bp or 99-bp repeat motif and are not shown). Statistical significance was calculated using a Wilcoxon test. Underlying data can be found in S1 Data. H3K9me3, trimethylation of histone 3 lysine 9.
(PDF)

S14 Fig. Validation of lack of chimeric sequence assemblies for neo-X and neo-Y regions. A. Shown is mapping of Illumina reads from neo-X-derived BAC clones to their (I) correct neo-X genomic location, (II) to their homologous neo-Y region, and (III) to their homologous neo-Y region after masking the correct neo-X location. We see little cross-mapping of neo-X-derived BAC clone reads to the homologous neo-Y location (see II), and neo-X reads only start mapping to their homologous neo-Y region (with many SNPs, as indicated by the colors in the coverage track) after the neo-X region is masked (see III), revealing their former

homology. Little cross-mapping of neo-X reads to the neo-Y chromosome confirms the high quality of our assembly, and lack of chimeric sequences. Also note that the homologous neo-Y segment is considerably larger than the neo-X, because of the accumulation of repetitive sequences on the neo-Y. **B.** Shown is mapping of Illumina reads from neo-Y-derived BAC clones to their (I) correct neo-Y genomic location, (II) to their homologous neo-X region, and (III) to their homologous neo-X region after masking the correct neo-Y location. We see little cross-mapping of neo-Y-derived BAC clone reads to their homologous neo-X location (see II), and neo-Y reads only start mapping to their homologous neo-X region (with many SNPs, as indicated by the colors in the coverage track) after the neo-Y region is masked (see III), revealing their former homology. Little cross-mapping of neo-Y reads to the neo-X chromosome confirms the high quality of our assembly, and lack of chimeric sequences. Also note that the homologous neo-X segment is considerably smaller than the neo-Y, because of the accumulation of repetitive sequences on the neo-Y. BAC, bacterial artificial chromosome.
(PDF)

S15 Fig. Neo-X versus neo-Y assembly. A. Homologous blocks on the neo-X and neo-Y. **B.** Global alignments between homologous regions on the neo-X and neo-Y. **C.** Local alignments between homologous neo-X and neo-Y regions chosen at random. The red bars indicate repeats. The scaffolds neo-Y1 and neo-Y2 refer to the two largest neo-Y scaffolds (see Fig 2C).
(PDF)

S16 Fig. Distribution of different repeat types across the *D. miranda* genome. The pie chart insert shows the relative types of different repeats across chromosomes, and the bar charts show the absolute number of bases masked for the various repeats across chromosomes. Underlying data can be found in S1 Data.
(PDF)

S17 Fig. Single-copy Y genes have been translocated onto the dot chromosome. A. Comparison between *D. pseudoobscura* and *D. miranda* dot chromosome reveals an inversion between species involving the translocated ancestral Y region. **B.** BAC clone S513-N704 spans the Y-dot translocation and contains both ancestral Y genes (*kl-3*) as well as genes from Muller element F (shown in orange). **C.** H3K9me3 enrichment at genes on the dot that are derived from the ancestral Y versus Muller element F. **D.** Ancestral Y genes show testis-specific expression. Underlying data can be found in S1 Data. BAC, bacterial artificial chromosome; H3K9me3, trimethylation of histone 3 lysine 9.
(PDF)

S18 Fig. Length distribution of PacBio reads. A total of 2,407,465 reads with an average length of 12,818 bp and NR50 of 17,116 bp were collected. NR50, read length such that 50% of the total sequence is contained within reads of this length or longer.
(PDF)

S19 Fig. BioNano data. A. Molecule length distribution, **B.** molecule length versus molecule average intensity.
(PDF)

S1 Table. Data used for the current assembly.
(PDF)

S2 Table. Manual corrections of assembly.
(PDF)

S3 Table. Structural variants identified using Lumpyexpress, by mapping two male MSH22 Illumina libraries back to the reference genomes (626-bp and 915-bp insert sizes). Note that the published *D. miranda* genome is substantially smaller and lacks an assembly of repeat-rich regions and the Y/neo-Y chromosome.
(PDF)

S4 Table. BioNano Data: hybrid scaffold coverage by BioNano contigs and our NGS contigs. NGS, next-generation sequencing.
(PDF)

S5 Table. Summary of mapping location of BAC clone data. BAC, bacterial artificial chromosome.
(PDF)

S6 Table. Mapping location of BAC clone data. BAC, bacterial artificial chromosome.
(PDF)

S7 Table. Comparison of current and previous assembly of *D. miranda*.
(PDF)

S8 Table. BUSCO analysis of assemblies. BUSCO, Benchmarking Universal Single-Copy Orthologs.
(PDF)

S9 Table. Molecule statistics for BioNano data. A total of 90,977 molecules were obtained.
(PDF)

S10 Table. cDNA libraries used for annotation.
(PDF)

S11 Table. Full list of accession numbers (see data availability statement).
(XLSX)

S1 Data.
(XLSX)

Acknowledgments

We thank J.J. Emerson, Mahul Chakraborty, Olga Dudchenko, Ryan Bracewell, Emily Brown, and Alison Nguyen for technical assistance.

Author Contributions

Conceptualization: Doris Bachtrog.

Data curation: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco.

Formal analysis: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco.

Funding acquisition: Doris Bachtrog.

Investigation: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco, Doris Bachtrog.

Methodology: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Doris Bachtrog.

Project administration: Doris Bachtrog.

Resources: Doris Bachtrog.

Supervision: Doris Bachtrog.

Validation: Shivani Mahajan, Doris Bachtrog.

Visualization: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco, Doris Bachtrog.

Writing – original draft: Doris Bachtrog.

Writing – review & editing: Shivani Mahajan, Kevin H.-C. Wei, Matthew J. Nalley, Lauren Gibilisco, Doris Bachtrog.

References

1. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, et al. Sex determination: why so many ways of doing it? PLoS Biol. Public Library of Science; 2014; 12: e1001899. https://doi.org/10.1371/journal.pbio.1001899 PMID: 24983465
2. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet. Nature Publishing Group; 2006; 7: 645–653. https://doi.org/10.1038/nrg1914 PMID: 16847464
3. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. Nature Publishing Group; 2013; 14: 113–124. https://doi.org/10.1038/nrg3366 PMID: 23329112
4. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res. 2017; 27: 709–721. https://doi.org/10.1101/gr.213512.116 PMID: 28373483
5. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. Cold Spring Harbor Lab; 2010; 20: 1165–1173. https://doi.org/10.1101/gr.101360.109 PMID: 20508146
6. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, et al. Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biol. BioMed Central; 2002; 3: RESEARCH0085. https://doi.org/10.1186/gb-2002-3-12-research0085 PMID: 12537574
7. Soh YQS, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell. 2014; 159: 800–813. https://doi.org/10.1016/j.cell.2014.09.052 PMID: 25417157
8. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. Nature Publishing Group; 2003; 423: 825–837. https://doi.org/10.1038/nature01722 PMID: 12815422
9. Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, et al. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat Genet. Nature Publishing Group; 2017; 49: 387–394. https://doi.org/10.1038/ng.3778 PMID: 28135246
10. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. Nature Publishing Group; 2010; 463: 536–539. https://doi.org/10.1038/nature08700 PMID: 20072128
11. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014; 508: 494–499. https://doi.org/10.1038/nature13206 PMID: 24759411
12. Lucchesi JC. Gene dosage compensation and the evolution of sex chromosomes. Science. 1978; 202: 711–716. PMID: 715437
13. Bachtrog D, Charlesworth B. Reduced adaptation of a non-recombining neo-Y chromosome. Nature. Nature Publishing Group; 2002; 416: 323–326. https://doi.org/10.1038/416323a PMID: 11907578
14. Zhou Q, Ellison CE, Kaiser VB, Alekseyenko AA, Gorchakov AA, Bachtrog D. The epigenome of evolving *Drosophila* neo-sex chromosomes: dosage compensation and heterochromatin formation. Becker PB, editor. PLoS Biol. Public Library of Science; 2013; 11: e1001711. https://doi.org/10.1371/journal.pbio.1001711 PMID: 24265597
15. Ellison CE, Bachtrog D. Dosage compensation via transposable element mediated rewiring of a regulatory network. Science. American Association for the Advancement of Science; 2013; 342: 846–850. https://doi.org/10.1126/science.1239552 PMID: 24233721

16. Carvalho AB, Clark AG. Y chromosome of *Drosophila pseudoobscura* is not homologous to the ancestral Drosophila Y. Science. American Association for the Advancement of Science; 2005; 307: 108–110. https://doi.org/10.1126/science.1101675 PMID: 15528405

17. Larracuente AM, Noor MAF, Clark AG. Translocation of Y-linked genes to the dot chromosome in *Drosophila pseudoobscura*. Mol Biol Evol. 2010; 27: 1612–1620. https://doi.org/10.1093/molbev/msq045 PMID: 20147437

18. Chang C-H, Larracuente AM. Genomic changes following the reversal of a Y chromosome to an autosome in *Drosophila pseudoobscura*. Evolution. 2017; 71: 1285–1296. https://doi.org/10.1111/evo.13229 PMID: 28322435

19. Bachtrog D, Hom E, Wong KM, Maside X, de Jong P. Genomic degradation of a young Y chromosome in *Drosophila miranda*. Genome Biol. BioMed Central; 2008; 9: R30. https://doi.org/10.1186/gb-2008-9-2-r30 PMID: 18269752

20. Zhou Q, Bachtrog D. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. Science. American Association for the Advancement of Science; 2012; 337: 341–345. https://doi.org/10.1126/science.1225385 PMID: 22822149

21. Kaiser VB, Bachtrog D. *De novo* transcriptome assembly reveals sex-specific selection acting on evolving neo-sex chromosomes in *Drosophila miranda*. BMC Genomics. BioMed Central; 2014; 15: 241. https://doi.org/10.1186/1471-2164-15-241 PMID: 24673816

22. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017; 431: 931. https://doi.org/10.1038/ng.3802

23. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. Nature Research; 2016; 13: 1050–1054. https://doi.org/10.1038/nmeth.4035 PMID: 27749838

24. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27: 722–736. https://doi.org/10.1101/gr.215087.116 PMID: 28298431

25. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016; 44: e147. https://doi.org/10.1093/nar/gkw654 PMID: 27458204

26. Chu C, Nielsen R, Wu Y. REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. Antoniewski C, editor. PLoS ONE. 2016; 11: e0150719. https://doi.org/10.1371/journal.pone.0150719 PMID: 26977803

27. Smith A, Hubley R. RepeatModeler Open-1.0. In: RepeatMasker Open-4.0.

28. Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2014; 48: 4.11.1–39. https://doi.org/10.1002/0471250953.bi0411s48 PMID: 25501943

29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

30. Bartolomé C, Charlesworth B. Rates and patterns of chromosomal evolution in *Drosophila pseudoobscura* and *D. miranda*. Genetics. Genetics; 2006; 173: 779–791. https://doi.org/10.1534/genetics.105.054585 PMID: 16547107

31. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. Cold Spring Harbor Lab; 2015; 25: 445–458. https://doi.org/10.1101/gr.185579.114 PMID: 25589440

32. Pardue M-L, DeBaryshe PG. Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci USA. 2011; 108: 20317–20324. https://doi.org/10.1073/pnas.1100278108 PMID: 21821789

33. Villasante A, Abad JP, Planelló R, Méndez-Lago M, Celniker SE, de Pablos B. *Drosophila* telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. Cold Spring Harbor Lab; 2007; 17: 1909–1918. https://doi.org/10.1101/gr.6365107 PMID: 17989257

34. Steinemann M, Nauber U. Frequency of telomere repeat units in the *Drosophila miranda* genome. Genetica. 1986; 69: 47–57. https://doi.org/10.1007/BF00122933

35. Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell. American Society of Plant Biologists; 2012; 24: 3559–3574. https://doi.org/10.1105/tpc.112.100511 PMID: 22968715

36. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. BioMed Central; 2013; 14: R10. https://doi.org/10.1186/gb-2013-14-1-r10 PMID: 23363705

37. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. American Association for the Advancement of Science; 2001; 293: 1098–1102. https://doi.org/10.1126/science.1062939 PMID: 11498581

38. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. Oxford University Press; 1999; 27: 573–580. PMID: 9862982

39. Tek AL, Kashihara K, Murata M, Nagaki K. Functional centromeres in *Astragalus sinicus* include a compact centromere-specific histone H3 and a 20-bp tandem repeat. Chromosome Res. Springer Netherlands; 2011; 19: 969–978. https://doi.org/10.1007/s10577-011-9247-y PMID: 22065151

40. Talbert P, Kasinathan S, Henikoff S. Simple and Complex Centromeric Satellites in *Drosophila* Sibling Species. Genetics. Genetics; 2018;: genetics.300620.2017. https://doi.org/10.1534/genetics.117.300620 PMID: 29305387

41. Steiner FA, Henikoff S. Diversity in the organization of centromeric chromatin. Curr Opin Genet Dev. 2015; 31: 28–35. https://doi.org/10.1016/j.gde.2015.03.010 PMID: 25956076

42. Steinemann M, Steinemann S. Degenerating Y chromosome of *Drosophila miranda*: a trap for retrotransposons. Proc Natl Acad Sci USA. National Academy of Sciences; 1992; 89: 7591–7595. PMID: 1323846

43. Charlesworth B. Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci USA. National Academy of Sciences; 1978; 75: 5618–5622. PMID: 281711

44. Mahajan S, Bachtrog D. Convergent evolution of Y chromosome gene content in flies. Nat Commun. Nature Publishing Group; 2017; 8: 785. https://doi.org/10.1038/s41467-017-00653-x PMID: 28978907

45. Muller HJ. Bearings of the Drosophila work on systematics. In: JS H, editor. The new Systematics. Oxford; 1940.

46. Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci USA. National Acad Sciences; 2001; 98: 13225–13230. https://doi.org/10.1073/pnas.231484998 PMID: 11687639

47. Koerich LB, Wang X, Clark AG, Carvalho AB. Low conservation of gene content in the *Drosophila* Y chromosome. Nature. Nature Publishing Group; 2008; 456: 949–951. https://doi.org/10.1038/nature07463 PMID: 19011613

48. McKee BD, Karpen GH. Drosophila ribosomal RNA genes function as an X-Y pairing site during male meiosis. Cell. 1990; 61: 61–72. PMID: 2156630

49. McKee BD, Habera L, Vrana JA. Evidence that intergenic spacer repeats of *Drosophila melanogaster* rRNA genes function as X-Y pairing sites in male meiosis, and a general model for achiasmatic pairing. Genetics. Genetics Society of America; 1992; 132: 529–544. PMID: 1330825

50. Ault JG, Rieder CL. Meiosis in Drosophila males. I. The question of separate conjunctive mechanisms for the XY and autosomal bivalents. Chromosoma. 1994; 103: 352–356. PMID: 7821091

51. Cooper KW. The mechanism of non-random segregation of sex chromosomes in male *Drosophila miranda*. Genetics. Genetics Society of America; 1946; 31: 181–194. PMID: 21021046

52. Roy V, Monti-Dedieu L, Chaminade N, Siljak-Yakovlev S, Aulard S, Lemeunier F, et al. Evolution of the chromosomal location of rDNA genes in two Drosophila species subgroups: ananassae and melanogaster. Heredity (Edinb). 2005; 94: 388–395. https://doi.org/10.1038/sj.hdy.6800612 PMID: 15726113

53. White MJD. Cytological Evidence on the Phylogeny and Classification of the Diptera. Evolution. 1949; 3: 252. https://doi.org/10.2307/2405562 PMID: 18138385

54. Steinemann M. Telomere repeats within the neo-Y-chromosome of *Drosophila miranda*. Chromosoma. 1984; 90: 1–5. https://doi.org/10.1007/BF00352271

55. Tomaszkiewicz M, Rangavittal S, Cechova M, Campos Sanchez R, Fescemyer HW, Harris R, et al. A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. Genome Res. 2016; 26: 530–540. https://doi.org/10.1101/gr.199448.115 PMID: 26934921

56. Hall AB, Papathanos P-A, Sharma A, Cheng C, Akbari OS, Assour L, et al. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. Proc Natl Acad Sci USA. National Acad Sciences; 2016; 113: E2114–23. https://doi.org/10.1073/pnas.1525164113 PMID: 27035980

57. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, et al. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol. 2018; 36: 321–323. https://doi.org/10.1038/nbt.4109 PMID: 29553574

58. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. Nature Publishing Group; 2013; 20: 267–273. https://doi.org/10.1038/nsmb.2506 PMID: 23463311

59. Prytkova TR, Zhu X, Widom J, Schatz GC. Modeling DNA-bending in the nucleosome: role of AA periodicity. J Phys Chem B. American Chemical Society; 2011; 115: 8638–8644. https://doi.org/10.1021/jp203564q PMID: 21639136

60. Ming R, Bendahmane A, Renner SS. Sex chromosomes in land plants. Annu Rev Plant Biol. 2011; 62: 485–514. https://doi.org/10.1146/annurev-arplant-042110-103914 PMID: 21526970

61. Muyle A, Shearn R, Marais GA. The Evolution of Sex Chromosomes and Dosage Compensation in Plants. Genome Biol Evol. 2017; 9: 627–645. https://doi.org/10.1093/gbe/evw282 PMID: 28391324

62. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. Nature Publishing Group; 2013; 10: 563–569. https://doi.org/10.1038/nmeth.2474 PMID: 23644548

63. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. Wang J, editor. PLoS ONE. Public Library of Science; 2014; 9: e112963. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509

64. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. American Association for the Advancement of Science; 2009; 326: 289–293. https://doi.org/10.1126/science.1181369 PMID: 19815776

65. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016; 3: 95–98. https://doi.org/10.1016/j.cels.2016.07.002 PMID: 27467249

66. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017; 356: 92–95. https://doi.org/10.1126/science.aal3327 PMID: 28336562

67. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

68. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010; 38: 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

69. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. Nature Publishing Group; 2012; 9: 357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

72. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. BioMed Central; 2014; 15: R84. https://doi.org/10.1186/gb-2014-15-6-r84 PMID: 24970577

73. Gramates LS, Marygold SJ, Santos GD, Urbano J-M, Antonazzo G, Matthews BB, et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 2017; 45: D663–D671. https://doi.org/10.1093/nar/gkw1016 PMID: 27799470

74. Smith A, Hubley R, Green P. RepeatMasker Open-4.0. In: RepeatMasker Open-4.0.

75. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. Nature Publishing Group; 2015; 12: 357–360. https://doi.org/10.1038/nmeth.3317 PMID: 25751142

76. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. Nature Publishing Group; 2015; 33: 290–295. https://doi.org/10.1038/nbt.3122 PMID: 25690850

77. Korf I. Gene finding in novel genomes. BMC Bioinformatics. BioMed Central; 2004; 5: 59. https://doi.org/10.1186/1471-2105-5-59 PMID: 15144565

78. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003; 19 Suppl 2: ii215–25.

79. Guizard S, Piégu B, Bigot Y. DensityMap: a genome viewer for illustrating the densities of features. BMC Bioinformatics. BioMed Central; 2016; 17: 204. https://doi.org/10.1186/s12859-016-1055-0 PMID: 27153821

80. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. BioMed Central; 2004; 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12 PMID: 14759262

81. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 2011; 39: e68–e68. https://doi.org/10.1093/nar/gkr123 PMID: 21398631

82. Noé L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 2005; 33: W540–3. https://doi.org/10.1093/nar/gki478 PMID: 15980530