Title

Sex Chromosome Evolution in Insects

Author

Mahajan, Shivani

Publication Date

2017

Sex Chromosome Evolution in Insects

By

Shivani Mahajan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor Doris Bachtrog, Chair
Professor Rasmus Nielsen
Professor Lior Pachter

Fall 2017

Abstract

Sex Chromosome Evolution in Insects

by

Shivani Mahajan

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Doris Bachtrog, Chair

Sex chromosomes play a role in sex determination in several organisms, ranging from humans and other mammals to flies, and present unique characteristics that distinguish them from autosomes. However, the underlying evolutionary forces that drive sex chromosome evolution and the molecular processes and mechanisms shaping their unusual characteristics are poorly understood, even in well-studied organisms like *Drosophila melanogaster* and humans. In this dissertation, I studied several aspects of sex chromosome evolution, including the mechanism of dosage compensation, Y degeneration and gene content evolution of Y chromosomes, sequence evolution of recently formed neo-sex chromosomes, and changes of the chromatin landscape in species with newly evolved sex chromosomes, in both model and non-model organisms.

X and Y chromosomes are derived from a pair of ordinary autosomes. Y chromosomes do not recombine, which leads to their degeneration and causes them to lose genes. This creates a gene-dose imbalance for X-linked genes in males compared to females, and also compared to autosomes, and different organisms have evolved different mechanisms to compensate for this imbalance. While in humans one of the two X chromosomes in females is randomly inactivated in different cells, *Drosophila melanogaster* males hyper-transcribe genes on their single X chromosome. In the first chapter, I tested for dosage compensation in the order Strepsiptera, which is a sister group of Coleoptera (beetles). Using DNA-seq and RNA-seq data, I showed that the species *Xenos vesparum* shares an X chromosome with the flour beetle *Tribolium* that is fully dosage compensated. However, *X. vesparum* also contains a more recently evolved X chromosome that is autosomal in *Tribolium*, and which has evolved only partial dosage compensation.

Y chromosomes degenerate and contain very few genes. They also accumulate repetitive DNA, which makes their sequencing and assembly extremely difficult. In the second chapter I developed a bioinformatics pipeline to extract Y-linked coding sequences using

DNA-seq and RNA-seq data from males and females of a species, without having to assemble the repetitive Y chromosome. I applied this pipeline to several Diptera flies, to characterize and study their Y gene content. I showed that there was no overlap between Y-linked genes in different Dipterans, and that different species had convergently acquired genes with testis-specific functions, highlighting the importance of male-specific selection in driving the evolution of Y gene content.

Species with newly evolved neo-sex chromosomes, such as *Drosophila miranda*, provide a unique opportunity to study sex chromosome evolution, since its neo-Y still retains significant sequence identity to its former homolog, the neo-X chromosome, and it also still contains thousands of genes. However, this also makes the neo-Y chromosome particularly difficult to assemble using short read technology alone, due to the inability to unambiguously assign sequencing reads to either the neo-X or the neo-Y, and also due to the repetitive nature of the neo-Y in general. In the third chapter, I used Single Molecule Sequencing (Pacbio) and Chromatin Conformation Capture along with Illumina whole genome shot-gun sequencing to build a high quality genome assembly for *Drosophila miranda.* I showed that the neo-Y chromosome has greatly increased in size by almost 3-fold, compared to the neo-X chromosome, due to the accumulation of repetitive sequences, but also due to the expansion of some male-specific genes on the neo-Y. This assembly provides the basis for future functional studies of sex chromosome evolution in this species.

A large proportion of the genome in *Drosophila miranda* is repetitive and heterochromatic (~43%). The different chromatin compartments are established during early embryonic development, but very little is known about how this happens at the molecular level, and what primary sequences target parts of the genome to establish a heterochromatic conformation. In the fourth chapter I studied heterochromatin establishment in *D.miranda* during early development using single embryo ChIP-seq. I showed that males experience a delay in the establishment of the heterochromatic histone mark H3K9me3 compared to females. I also investigated signatures of H3K9me3 spreading near euchromatic transposable element (TE)/repeat insertions and showed that this signal is more pronounced for TEs that are targeted by maternally inherited piRNAs, suggesting that they may play an important role in the establishment of heterochromatin.

Dedication

To my Mother and Father
Upasana Mahajan
Rajiv Kumar Mahajan

&

To my PhD Advisor
Professor Doris Bachtrog

Acknowledgements

I would like to thank my PhD Advisor Dr. Doris Bachtrog for her constant guidance and support. I would also like to thank my lab members, both former and current, who have provided me with invaluable help and advice throughout my PhD. I would especially like to thank Dr. Beatriz Vicoso, a former lab member for being my mentor in my initial years as a PhD student. I would also like to thank Carolus Chan for his help with data collection and sequencing for my final chapter. I would like to thank my committee members, Rasmus Nielsen and Lior Pachter, for their guidance and feedback.

I would like to especially thank my labmates Dat Mai and Lauren Gibilisco for their friendship and for providing me with emotional support during these past five years.

I would also like to thank my boyfriend Pierre Masse for his patience, love and support during these past few months while I was writing my thesis.

Finally, I would like to thank my parents for their love and support and for always encouraging me to pursue my dreams.

Chapter 1

Partial dosage compensation in Strepsiptera, a sister group of beetles

Shivani Mahajan & Doris Bachtrog

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

Abstract

Sex chromosomes have evolved independently in many different taxa, and so have mechanisms to compensate for expression differences on sex chromosomes in males and females. Different clades have evolved vastly different ways to achieve dosage compensation, including hyper-transcription of the single X in male Drosophila, down-regulation of both X's in XX *Caenorhabditis*, or inactivation of one X in female mammals. In the flour beetle *Tribolium*, the X appears hyper-expressed in both sexes, which might represent the first of two steps to evolve dosage compensation along the paths mammals may have taken (i.e. up-regulation of X in both sexes, followed by inactivation of one X in females). Here we test for dosage compensation in Strepsiptera, a sister taxon to beetles. We identify sex-linked chromosomes in *Xenos vesparum* based on genomic analysis of males and females, and show that its sex chromosome consists of two chromosomal arms in *Tribolium*: the X chromosome that is shared between *Tribolium* and Strepsiptera, and another chromosome that is autosomal in *Tribolium* and another distantly related Strepsiptera species, but sex-linked in *X. vesparum*. We use RNA-seq to show that dosage compensation along the X of *X. vesparum* is partial and heterogeneous. In particular, genes that are X-linked in both beetles and Strepsiptera appear fully dosage compensated probably through down-regulation in both sexes, while genes on the more recently added X segment have evolved only partial dosage compensation. In addition, reanalysis of published RNA-seq data suggests that *Tribolium* has evolved dosage compensation, without hypertranscribing the X in females. Our results demonstrate that patterns of dosage compensation are highly variable across sex-determination systems and even within species.

Introduction

Heteromorphic sex chromosomes have arisen independently in many species from ordinary autosomes (Bull 1983; Charlesworth 1996). Sex chromosome evolution is characterized by a loss of gene function along the non-recombining Y chromosome (Y degeneration; see Bachtrog (2013) for a recent review). In many organisms with heteromorphic XY sex chromosomes, mechanisms have evolved that equalize expression of X-linked genes in males and females (dosage compensation; Charlesworth 1996; Vicoso and Bachtrog 2009).

Dosage compensation has evolved in response to reduced gene dose of X-linked genes in males, due to loss of Y-linked genes (Charlesworth 1996; Vicoso and Bachtrog 2009). The most direct way to achieve dosage compensation is to simply up-regulate X-linked genes in males only, to restore correct levels of X-linked gene product in males, as has evolved in Drosophila (Gelbart and Kuroda 2009). However, in the other model systems where dosage compensation has been well studied - mammals and *Caenorhabditis* - the dosage compensation mechanisms operate by reducing expression of the X chromosome in XX females (or hermaphrodites), with mammals completely inactivating one of the two X's in females, and *Caenorhabditis* halving expression of each X in a hermaphrodite (Heard and Disteche 2006; Meyer 2000).

Halving expression of the X in females presents somewhat of an evolutionary conundrum. If dosage compensation evolved to counterbalance reduced expression of X-linked genes in males in response to Y degeneration and to restore the correct balance between X-linked and autosomal gene products in males, the down-regulation of gene expression on the X in females does not solve the gene dose problem that males experience. Instead, it simply creates the same gene dose deficiency and X-autosome imbalances of gene products in females. It has thus been proposed that dosage compensation in mammals and *Caenorhabditis* evolved in a two-step process (Charlesworth 1996; Mank 2013; Mank, et al. 2011; Ohno 1967; Vicoso and Bachtrog 2009). In response to Y degeneration, the X first became up-regulated in both sexes. This would have resolved the gene dose deficiency that is experienced by males, but would also result in too much gene product in females. In response to over-expression in females, X down-regulation or X inactivation has evolved secondarily, to restore correct X-autosome gene balance in females (Charlesworth 1996; Ohno 1967; Vicoso and Bachtrog 2009).

In both mammals and *Caenorhabditis*, only the second step of dosage compensation is well understood, and there has been considerable debate on whether there is global up-regulation of X-linked genes relative to its ancestral expression level (Nguyen and Disteche 2006). Depending on the data set used, and statistical analysis of sometimes the same data, different studies have yielded opposite conclusions as to whether mammals globally up-regulate their X chromosome in both sexes, relative to its ancestral expression level, or relative to autosomal expression, or whether only a subset of dosage-sensitive genes is up-regulated on the X (Pessia, et al. 2012). Most recent studies have concluded that placental mammals do not globally up-regulate their X chromosome, but instead, only a subset of X-linked genes that are part of protein-complexes appear to be up-regulated in both sexes (Lin, et al. 2012; Pessia, et al. 2012), while a subset of autosomal genes interacting with X-linked genes were found to have become down-regulated in placentals upon the emergence of sex chromosomes (Julien, et al. 2012). Marsupials, on the other hand, may globally upregulate their X in both sexes, relative to ancestral expression levels (Julien, et al. 2012). Thus, the mechanisms of dosage compensation, and the evolution of X inactivation in mammals remains controversial (Julien, et al. 2012; Pessia, et al. 2013). Studying additional taxa with independently formed sex chromosomes should help to identify general principles driving the evolution of dosage compensation.

Karyotypes in Coleoptera have been well-studied, and almost all beetles have heteromorphic sex chromosomes (either XX/XY or XX/X0 systems). A recent study in the flour beetle *Tribolium castaneum*, a model species whose genome has been sequenced and annotated, has concluded that dosage compensation in this species evolved by up-regulating the X chromosome in a non-sex-specific manner, i.e. expression of the X was increased in both males and females (Prince, et al. 2010). While this restores correct expression of X-linked genes in males, it also leads to hyper-transcription of the X in females (Prince, et al. 2010), and may thus represent the first of two steps to evolve dosage compensation along the paths marsupials may have taken. Here we identify sex-linked genes and analyze male and female gene expression in twisted-wing insects (Strepsiptera), a sister taxon to beetles (Coleoptera), to better understand the evolutionary forces driving dosage compensation in this group.

Strepsiptera are a morphologically highly derived group of endoparastic insects whose phylogenetic position was debated for some time, but the most recent and complete studies clearly support a sister relationship of Strepsiptera with Coleoptera (Boussau, et al. 2014; Niehuis, et al. 2012). Strepsiptera have separate sexes, and cytogenetic data for this group exists for just two species but indicates the presence of heteromorphic X and Y chromosomes. In particular, the diploid chromosome number of *Xenos peckii* was identified as 16, and in an unidentified species of *Xenos* from Brazil, three pairs of autosomes and an XY sex chromosome were reported (Ferreira et al. 1984). Here we use genomic sequencing of the Strepsiptera *Xenos vesparum* (family Stylopidae), and we also analyze published genome data from *Mengenilla moldrzyki* (Niehuis, et al. 2012), a species belonging to the early-divergent Strepsipteran family Mengenillidae, to identify the sex chromosomes of Strepsiptera, and gene expression analysis in *X. vesparum* and *T. castaneum* to investigate the absence or presence of dosage compensation.

Materials & Methods

Sampling and sequencing of Strepsiptera

We sequenced the DNA from an adult male (library insert size 700-800 bp) and two females of *X. vesparum* (neotenic adult female with library insert size 700-800 bp and female 4th instar larva with library insert size of 250 bp). For gene expression analysis, we prepared libraries for two female samples (neotenic adults, and 4th instar larvae; library insert size about 200 bp), and one male sample (pupae). DNA was extracted using Puregene, with proteinase K and RNAse A treatment during lysis, and was purified with overnight Isopropanol precipitation. RNA was extracted with Trizol, and purified overnight with Ethanol precipitation. For both the DNA and RNA extraction, purity measurement and quantification was done using Nanodrop and Qubit. The Libraries were prepared using standard Illumina TruSeq kits and protocols, and the cleanup was done using AmpureXP, followed by size-selection of the DNA libraries on agarose gels. We obtained 27,578,418 genomic reads for the adult female; 77,729,238 reads for male; and 19,045,611 reads for female larva, respectively. After RNA-seq we obtained 100,160,332 reads for the neotenic adult female; 314,698,728 reads for male; and 283,804,476 reads for female 4th instar larva, respectively. The genome assembly of *Mengenilla moldrzyki* was obtained from http://datadryad.org/resource/doi:10.5061/dryad.ts058.2, and unpaired shotgun 454 reads (a total of 5,449,680 reads) from male *M. moldrzyki* samples were provided to us by the authors (Niehuis, et al. 2012).

Genome Assembly & coverage analysis to infer sex-linkage

Paired-end reads from the female *X. vesparum* sample were trimmed and assembled using SOAPdenovo (Li, et al. 2009) with a K-mer size of 63. Gapcloser was used to further improve the quality of the assembly. The assembled genome contained 11,895 scaffolds and was 81.4-Mb long (**Table S1**). Only scaffolds >1000-bp were retained for further analysis. Coding sequences (CDS) from *Tribolium* were used to assign the scaffolds to chromosomes. *Tribolium* CDS were downloaded from ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/ (version 3) and mapped to the *X. vesparum* genome using blat with a translated database and query and only the best hit was kept. A scaffold was assigned to the consensus *Tribolium* chromosome in case more than one gene mapped to that scaffold; when only one gene mapped to a scaffold, it was assigned to the chromosome on which that gene was located. A total of 2,291 scaffolds mapped to *T. castaneum* chromosomes and only these were retained for further analysis. Male and female *X. vesparum* trimmed paired-end genomic reads were aligned separately to the *de novo* assembled *X. vesparum* genome using bwa (Li and Durbin 2009). Scaffold coverage was calculated using soapcoverage. The log (base 2) of the coverage per chromosome was then plotted in R. 454 reads from male *M. moldrzyki* samples were aligned to the published reference *M. moldrzyki* genome (Niehuis, et al. 2012) using bwa-sw and coverage was calculated using soapcoverage. Sex-linkage of scaffolds was inferred in the same way as for *X. vesparum* using *T. castaneum* coding sequences. Coverage was normalized by the median scaffold length as well as the median coverage of the autosomes, and log (base 2) of the normalized male coverage was plotted in R.

Transcriptome assembly & gene expression analysis

FastQC was used for the quality control of the paired end reads from the two female (4th instar larva and neotenic adult) samples and male (adult) sample. The reads were then trimmed, pooled and assembled using SOAPdenovotrans with a kmer size of 75 (**Table S2**). The obtained transcripts were mapped to *Tribolium* CDS using Blat with a translated query and database. The Blat output was then filtered and only the best match per transcript was retained. For transcripts overlapping a *Tribolium* gene by more than 20-bp, only the transcript with the highest alignment score was retained. For those that overlapped by less than 20-bp, their sequences were concatenated. Transcripts mapping to different parts of the same gene were also concatenated. Finally, transcripts were assigned the location of their corresponding *Tribolium* genes on the *Tribolium* genome. This resulted in a total of 4413 genes for *X. vesparum*. Trimmed male and female paired-end RNA-seq reads were aligned to the *de novo* assembled transcriptome using bowtie2 (Langmead and Salzberg 2012) and FPKM values were calculated using eXpress (Roberts and Pachter 2013). The log (base 2) of the FPKM values per chromosome were then plotted in R. We also analyzed published RNA-seq reads from *T. castaneum* (Li, et al. 2013). Unpaired RNA-seq reads from male and female abdominal and prothoracic glands were downloaded from NCBI SRA (http://www.ncbi.nlm.nih.gov/sra/; accession numbers: SRX501821, SRX501822, SRX501819 and SRX501820). For each gland, male and female reads were separately mapped to *Tribolium* CDS using bowtie2 and FPKM values were calculated using the eXpress package and plotted in R.

Results

Identification of sex chromosomes in Strepsiptera

To infer sex chromosomes in *Xenos*, we used genomic read coverage in males *vs*. females (Vicoso and Bachtrog 2013). In particular, regions that are autosomal should show equal read coverage in both sexes, while X-linked regions only have half the coverage in males relative to females. Indeed, we find a bimodal distribution of male/female coverage of scaffolds, indicating that a substantial fraction of the genomic scaffolds are X-linked in *Xenos* (**Figure 1A**). To order scaffolds from Strepsiptera, we mapped them against chromosomes from *Tribolium*. *T. castaneum* contains 10 similarly sized chromosome pairs, one of which (chromosome 1) segregates as the X chromosome. We find that scaffolds mapping to two chromosome elements from *Tribolium* show reduced male/female read coverage in *X. vesparum*, suggesting that the sex chromosomes of *Xenos* correspond to two different chromosomes of *Tribolium* (**Figure 1B,C; Fig. S1**). One of the X-linked elements of *X. vesparum* is also the X chromosome of *Tribolium*, which suggests that this chromosome may already have been a sex chromosome in an ancestor of beetles and Strepsiptera, and thus may have been segregating as a sex chromosome for over 250 MY (Wiegmann, et al. 2009). The other sex-linked element of *X. vesparum* corresponds to chromosome 4 of *Tribolium*, suggesting that this element became X-linked only after the split of Coleoptera and Strepsiptera. Coverage analysis of genomic reads from male *Mengenilla moldrzyki* (Niehuis, et al. 2012), a species belonging to the early-divergent strepsipteran family Mengenillidae, shows that only scaffolds that map to chromosome 1 of *Tribolium* have reduced read coverage in male *M. moldrzyki* (**Fig. S2**). This supports that chromosome 1 is an ancient sex chromosome in Strepsiptera, and also shows that chromosome 4 only became X-linked in an ancestor of *Xenos* after the divergence of the two families Mengenilidae and Stylopidae about 50 MY ago (Wiegmann, et al. 2009). We refer to these segments as the ancestral region (homologous to chromosome 1 of *Tribolium*) and more recently added region (homologous to chromosome 4 of *Tribolium*) of the X chromosome of *Xenos*.

To investigate if segments of other chromosomes also show reduced coverage, and if coverage along chromosome 1 and 4 is reduced uniformly, we mapped our *Xenos* scaffolds along the *Tribolium* genome (**Figure 2, Fig. S3, S4**). In general, we find no evidence of large genomic segments from other chromosomes to show reduced coverage in males vs. females (**Figure 2A**). Thus, while it is certainly the case that individual genes from these other chromosomes are also sex-linked in *Xenos*, most of them appear to be indeed autosomal in *X. vesparum.* On the other hand, coverage along chromosome 1 and 4 is reduced relatively uniformly (**Figure 2B**), suggesting that most genes located on these chromosomes are sex-linked in *Xenos*. Nevertheless, some scaffolds on chromosomes 1 and 4 clearly show coverage levels that suggest that they are autosomal in *Xenos* (**Figure 1C**), indicating that some genomic rearrangements have taken place between these species. We therefore used two approaches to identify X-linked and autosomal genes in *Xenos*: (1) Genes were considered X-linked if their reciprocal-best-hit in *Tribolium* was located on chromosome 1 or chromosome 4. (2) Genes were classified as X-linked if they were located on a scaffold that had reduced male/female coverage, as shown on **Figure 1A**: scaffolds in the green shaded area were classified as X-linked, scaffolds in the orange shaded area were classified as autosomal. Both classifications were used to compare the male and female expression of genes on the X and autosomes of *Xenos*, and both yielded similar results (see below).

Gene expression analysis in Strepsiptera and *Tribolium*

To assay if *Xenos* has evolved dosage compensation, we measured gene expression in males and females. FPKM cutoffs for each sample were determined based on FPKM values for introns and intergenic regions (see **Fig. S5**; note that the results are insensitive to different FPKM cutoffs, **Figs. S6, S7**). We assayed gene expression in neotenic adult females, 4th instar female larvae and male pupae of *X. vesparum* (**Figure 3**). Expression levels are similar across autosomes in both sexes, and similar to expression levels of genes mapping to *Tribolium* chromosome 4 (the more recently formed X) in females, but reduced in males (Wilcoxon test p-value 1.2e-07 when comparing expression of chromosome 4 in males versus autosomes; and p-values 3.0e-07 and < 2.2e-16 when comparing male/neotenic adult female and male/larval female FPKM ratios, respectively, for chromosome 4 versus the autosomes; **Figure 4A**). Expression from genes mapping to chromosome 1 of *Tribolium* (the ancestral X) is slightly reduced in both sexes, to a similar extent, relative to autosomes (Wilcoxon test p-values 0.0001, 0.0002 and 0.0256 for male, neotenic adult female and female larva, respectively; **Figure 4A**). Male/female expression ratios for both adult and larval sample are similar across autosomes, and X-linked genes mapping to chromosome 1 of *Tribolium*. Thus, genes on the presumably ancestral X of Strepsiptera and beetles are dosage compensated. In contrast, *X. vesparum* genes mapping to *Tribolium* chromosome 4 show significantly lower male/female expression ratios, using both the adult and larvae female sample (**Figure 3**). The decrease in expression in males is less than 0.5, suggesting that this more recently formed X chromosome has evolved partial but incomplete dosage compensation. Thus, we find nearly complete dosage compensation on the ancestral X, which is expressed at a lower level in both sexes relative to autosomes, and partial dosage compensation on the more recently added X, which shows lower expression relative to autosomes in males only.

We used expression in *Tribolium* as a proxy to infer ancestral expression levels in *Xenos*. A previous analysis of whole-body adult microarray data showed that *Tribolium* males have similar levels of gene expression at X-linked and autosomal genes, while the X appears hyper-transcribed in females (Prince, et al. 2010). We analyzed published RNA-seq data from prothoracic glands and abdominal glands from male and female *T. castaneum* and, surprisingly, found that expression of genes from the X and the autosomes is similar in both males and females, for both tissues (**Figure 4A, Fig. S8, S9**). This would suggest that *Tribolium*, at least in its prothoracic and abdominal glands, has evolved dosage compensation without hyper-transcribing the X chromosome in females. To test if dosage compensation in *Xenos* evolved through down-regulation of the X in both sexes (as has been suggested in mammals), or through up-regulation of the single X in males (as done in *Drosophila*), we used expression from *Tribolium* as a proxy for proto-X expression before the X became sex-linked. Contrasting expression levels of the X on *Xenos* to that of *Tribolium* suggests that genes mapping to chromosome 1 have become down-regulated in both male and female *Xenos*, while genes mapping to chromosome 4 are expressed at a lower level only in male *Xenos* (**Figure 4B, Fig. S10**). Thus, this supports our conclusion that the ancestral X chromosome of *Xenos* is dosage compensated through down-regulation in both sexes, and only partial dosage compensation has evolved on the more recently formed X chromosome. Note that this analysis assumes that gene expression levels on chromosome 1 in *Tribolium*, which is X-linked in both species, reflect ancestral expression levels. As stated above, expression from the X is similar in both male and female *Tribolium*, and similar to autosomes. If expression on the X in females reflects ancestral expression levels in *Tribolium* and dosage compensation simply evolved by up-regulating the X in males (as done in Drosophila), this would validate the use of *Tribolium* expression data from chromosome 1 as a proxy for ancestral expression levels. It is also possible that expression on chromosome 1 was higher ancestrally, and dosage compensation in *Tribolium* evolved by down-regulating the X in both sexes (as has happened in mammals); this would imply that we underestimate the magnitude of down-regulation on the *Xenos* X chromosome. However, we cannot exclude the formal possibility that chromosome 1 was expressed at a lower level in an ancestor of beetles and became up-regulated in both male and female *Tribolium* after it became a sex chromosome but more so in males (to compensate for gene dose differences between sexes). In this case, expression from chromosome 1 may not have changed in female *Xenos*. Genes on chromosome 4 are autosomal in *Tribolium* and thus should reflect the ancestral proto-X expression level.

To investigate if dosage compensation is heterogeneous across the X chromosomes with some segments being compensated, we mapped male and female expression levels along

the *Tribolium* genome (**Figure 5**). We find that expression of genes mapping along chromosome 4 is generally female-biased, both in neotenic adult females, and 4th instar female larvae. This is consistent with a global lack of chromosome-wide dosage compensation along this chromosome. On the other hand, there is more heterogeneity in expression levels of genes mapping across chromosome 1, with some regions showing male-biased expression, and others showing female-biased expression, resulting in global patterns of dosage compensation on this chromosomal element (**Figure 4B**). This is similar to patterns of sex-biased gene expression seen on autosomes (**Fig. S11, S12**).
We repeated the gene expression analysis using only scaffolds that show reduced genomic coverage in males relative to females, and find similar results (**Fig. S13**). X-linked genes are generally under-expressed in males relative to autosomes, both for genes mapping to the ancestral and the more recently formed X, while in females, only genes on the ancestral X are down-regulated, relative to autosomes. Thus, the X of Strepsiptera shows partial dosage compensation, and genes mapping to chromosome 1 are compensated more fully than those mapping to chromosome 4. However, dosage compensation on the ancestral X seems less complete using this classification scheme, possibly due to inclusion of some autosomal genes with equal expression in males and females using the first classification method.

Discussion

Both beetles and Strepsiptera have heteromorphic sex chromosomes, and we show that the X chromosome of *X. vesparum* consists of two chromosomal elements in *Tribolium*. Part of the X chromosome of *X. vesparum* is homologous to the X of *Tribolium*, suggesting that this element was already an X chromosome in the ancestor of beetles and Strepsiptera, and thus has been segregating as a sex chromosome since before these groups split >250 MY ago. A second chromosome that is X-linked in *X. vesparum* is autosomal in both *Tribolium* and in a Strepsiptera species belonging to the early-divergent Strepsipteran family Mengenillidae, which implies that this chromosome became sex-linked more recently, after the split of Mengenilidae and Stylopidae over 50 MY ago. Without sampling of additional species, we cannot determine how long ago this second chromosome became incorporated into the X of *X. vesparum*. In species where chromosomes only became sex-linked relatively recently (i.e. in the past 1 MY), such as the neo-sex chromosomes of several Drosophila species, the X and the Y chromosome still harbor sufficient homology so that some sequencing reads that are derived from the Y chromosome also map to the X chromosome, and genomic coverage is only somewhat reduced for these recently formed neo-sex chromosomes (Vicoso and Bachtrog 2013; Zhou and Bachtrog 2012). Coverage of the more recently formed sex chromosome of *Xenos* is similar to that of the X shared with beetles, which suggests that this segment became X-linked long enough ago for its former homolog (the Y chromosome) to degenerate completely.

Whole-body microarray data suggested that dosage compensation in *Tribolium* involves the up-regulation of the X in both sexes (Prince, et al. 2010), resulting in female-biased expression of the X. Our analysis of RNA-seq data from prothoracic and abdominal glands, however, indicates that the X and autosomes are transcribed at similar levels in both sexes, i.e. dosage compensation has evolved without hyper-transcribing the X chromosome in females. It is possible that the different findings are due to differences in methodology, statistical analysis, or the mechanism of dosage compensation or sex-biased expression patterns among tissues. In particular, gonads of many organisms often show an excess of genes with sex-biased expression; Drosophila species, for example, often harbor an excess of genes with ovary-biased expression on their X chromosomes (Assis, et al. 2012), which could contribute to an excess of X-linked expression in whole-body adult females. On the other hand, testis may lack dosage compensation, as found in Drosophila (Rastelli and Kuroda 1998), reducing X-linked expression in whole body adult males. It will be of great interest to study gene expression in additional tissues of *Tribolium* as well as *Xenos*, to establish the mechanisms of dosage compensation and sex-biased expression patterns in these species, and how they vary across tissues.

We find that the two X-linked arms of *Xenos* show different levels of dosage compensation; the X shared between *Tribolium* and Strepsiptera appears to be expressed at a lower level in both male and female *X. vesparum*. Given that RNA-seq data suggest that the X chromosome is expressed at similar levels in male and female *Tribolium*, this probably does not reflect lower ancestral expression of that chromosomal arm, but instead suggests that genes mapping to chromosome 1 became down-regulated in both sexes of *Xenos*. Note that we do not have a suitable outgroup species where chromosome 1 is autosomal, so we cannot formally exclude the possibility that this chromosome was expressed at a lower level in an ancestor of beetles and Strepsiptera. Down-regulation of the X in females alone, however, does not restore gene dose imbalances between X-linked and autosomal genes, and dosage compensation of the ancestral X in Strepsiptera might involve the down-regulation of autosomal genes that interact with genes on the X, and evolve along the following path: Y degeneration creates gene dose imbalances for some networks that utilize both X-linked and autosomal genes in males, and down-regulation of autosomal genes that interact with X-linked genes would restore proper X-autosome expression ratios in males. If down-regulation of autosomal genes is not sex-specific, this would result in gene dose imbalances for these networks in females, and create selective pressure to down-regulate X-linked genes interacting with those autosomal genes in females. The outcome of this evolutionary process would be an X chromosome that is expressed at a lower level in both sexes relative to its ancestral expression level, and the simultaneous down-regulation of autosomal genes interacting with X-linked genes, in both sexes.

This path resembles the different proposed mechanisms of dosage compensation of the partially homologous X chromosomes shared by placental mammals and marsupials (Julien, et al. 2012; Pessia, et al. 2012). Gene expression analyses suggest that the X has become globally up-regulated in marsupials followed by X inactivation (Julien, et al. 2012), whereas no global up-regulation of the X chromosome was found in placental mammals. Instead, a subset of autosomal genes interacting with X-linked genes have become down-regulated (Julien, et al. 2012), and a subset of X-linked genes that are part of protein-complexes appear to have become up-regulated in placentals in both sexes upon the emergence of sex chromosomes (Lin, et al. 2012; Pessia, et al. 2012). Thus, different solutions were found to equilibrate X expression levels between the sexes in these two lineages, similar to what we find in Strepsiptera and Coleoptera.

Yet, while dosage compensation has evolved on the chromosomal arm that is also X-linked in *Tribolium*, dosage compensation appears incomplete at genes that locate to the more recently added part of the *X. vesparum* X chromosome. Several species with female heterogametic sex determination, including birds (Ellegren, et al. 2007), some butterflies (Harrison, et al. 2012) and snakes (Vicoso, et al. 2013), but also male heterogametic monotremes (Julien, et al. 2012), have not evolved chromosome-wide mechanisms to equalize X-linked expression levels in males and females. Incomplete dosage compensation implies that many sex-linked genes have different expression levels in males and females, and gene networks employing X-linked and autosomal genes will differ between sexes (Mank 2013). It is possible that there simply has not been enough time yet for this more recently formed X chromosome to evolve full dosage compensation, as has been proposed for the recently formed X chromosome of threespine sticklebacks (Leder, et al. 2010). It will be of great interest to study gene expression patterns in additional species of Coleoptera and Strepsiptera, to identify the strikingly different ways in which dosage alterations associated with the emergence of sex chromosomes were resolved.

Acknowledgements

Funded by NIH grants (R01GM076007 and R01GM093182) to D.B.

References

Assis R, Zhou Q, Bachtrog D 2012. Sex-biased transcriptome evolution in Drosophila. Genome Biol Evol 4:1189-1200.

Bachtrog D 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14: 113-124.

Boussau B, Walton Z, Delgado JA, Collantes F, Beani L, Stewart IJ, Cameron SA, Whitfield JB, Johnston JS, Holland PWH, Bachtrog D, Kathirithamby J, P. J, Huelsenbeck JP 2014. Strepsiptera, phylogenomics and the long branch attraction problem. 9:e107709.

Bull JJ. 1983. Evolution of Sex Determining Mechanisms. Menlo Park, CA: Benjamin Cummings.

Charlesworth B 1996. The evolution of chromosomal sex determination and dosage compensation. Curr. Biol. 6: 149-162.

Ellegren H, Hultin-Rosenberg L, Brunstrom B, Dencker L, Kultima K, Scholz B 2007. Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. BMC Biol 5: 40.

Ferreira A, Cella D, Mesa A, Virkki N (1984) Cytology and systematical position of Stylopids (Strepsiptera). Hereditas 100: 51-52.

Gelbart M, Kuroda M 2009. Drosophila dosage compensation: a complex voyage to the X chromosome. Development 136: 1399-1410.

Harrison PW, Mank JE, Wedell N 2012. Incomplete sex chromosome dosage compensation in the Indian meal moth, Plodia interpunctella, based on de novo transcriptome assembly. Genome Biol Evol 4: 1118-1126.

Heard E, Disteche CM 2006. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. Genes Dev 20: 1848 - 1867.

Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schutz F, Daish T, Grutzner F, Kaessmann H 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol 10: e1001328.

Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

Leder EH, Cano JM, Leinonen T, O'Hara RB, Nikinmaa M, Primmer CR, Merila J 2010. Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. Mol Biol Evol 27: 1495-1503.

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li J, Lehmann S, Weissbecker B, Ojeda Naharros I, Schutz S, Joop G, Wimmer EA 2013. Odoriferous Defensive stink gland transcriptome to identify novel genes necessary for quinone synthesis in the red flour beetle, Tribolium castaneum. PLoS Genet 9: e1003596.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967. Lin F, Xing K, Zhang J, He X 2012. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. Proc Natl Acad Sci U S A 109: 11752-11757.

Mank JE 2013. Sex chromosome dosage compensation: definitely not for everyone. Trends Genet 29: 677-683.

Mank JE, Hosken DJ, Wedell N 2011. Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. Evolution 65: 2133-2144. Meyer B 2000. Sex in the worm: counting and compensating X-chromosome dose. Trends Genet 16: 247 - 253.

Nguyen D, Disteche C 2006. Dosage compensation of the active X chromosome in mammals. Nat Genet 38: 47 - 53.

Niehuis O, Hartig G, Grath S, Pohl H, Lehmann J, Tafer H, Donath A, Krauss V, Eisenhardt C, J H, Petersen M, Mayer C, Meusemann K, Peters R, Stadler P, Beutel R, Bornberg-Bauer E, McKenna D, Misof B 2012. Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. Curr Biol 22: 1309-1313.

Ohno S. 1967. Sex chromosomes and sex linked genes. Berlin: Springer Verlag.

Pessia E, Engelstadter J, Marais GA 2013. The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? Cell Mol Life Sci. 1499-6

Pessia E, Makino T, Bailly-Bechet M, McLysaght A, Marais GA 2012. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. Proc Natl Acad Sci U S A 109: 5346-5351.

Prince EG, Kirkland D, Demuth JP 2010. Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. Genome Biol Evol 2: 336-346.

Rastelli L, Kuroda MI 1998. An analysis of maleless and histone H4 acetylation in Drosophila melanogaster spermatogenesis. Mech Dev 71: 107 - 117.

Roberts A, Pachter L 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10: 71-73.

Vicoso B, Bachtrog D 2009. Progress and prospects toward our understanding of the evolution of dosage compensation. Chromosome Research 17: 585-602.

Vicoso B, Bachtrog D 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. Nature 499: 332-335.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. PLoS Biology, 11:e1001643.

Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, Winterton SL, Yeates DK 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol 7: 34.

Zhou Q, Bachtrog D 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. Science 337: 341-345.

Figures

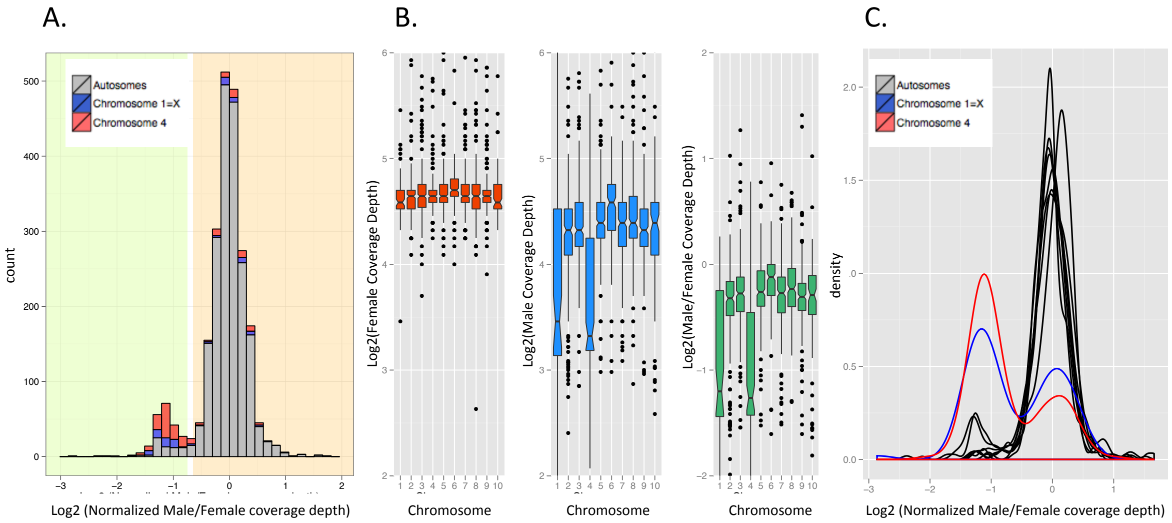


Figure 1 – Male and female genomic coverage analyses to identify sex chromosomes in *Xenos*. (A) Histogram of Log2 male/female coverage. Scaffolds that map to chromosomes 1 and 4 of *T. castaneum* are shown in blue and red, respectively. The bimodal distribution in coverage suggests that a substantial fraction of the genome is sex-linked in *Xenos*, with the peak with reduced male/female coverage corresponding to scaffolds that are X-linked in *X. vesparum*. (B) Boxplot of Log2 of coverage in female (in red), male (in blue), and male/female (in green). Overall, there is a drop in male/female coverage for scaffolds that map to chromosome 1 and 4 in *Tribolium*, suggesting that these chromosomal elements are X-linked in *X. vesparum*. (C) Density plot of log2 normalized male/female coverage. Normalization was done by dividing the coverage of scaffolds in each chromosome by the median of the coverage of all scaffolds in chromosomes other than chromosome 1 and chromosome 4. Distributions of chromosome 1 and chromosome 4 are different from that of other chromosomes. The bimodal shape suggests that there have been some rearrangements in the *X. vesparum* genome compared to the genome of T. *castaneum*.

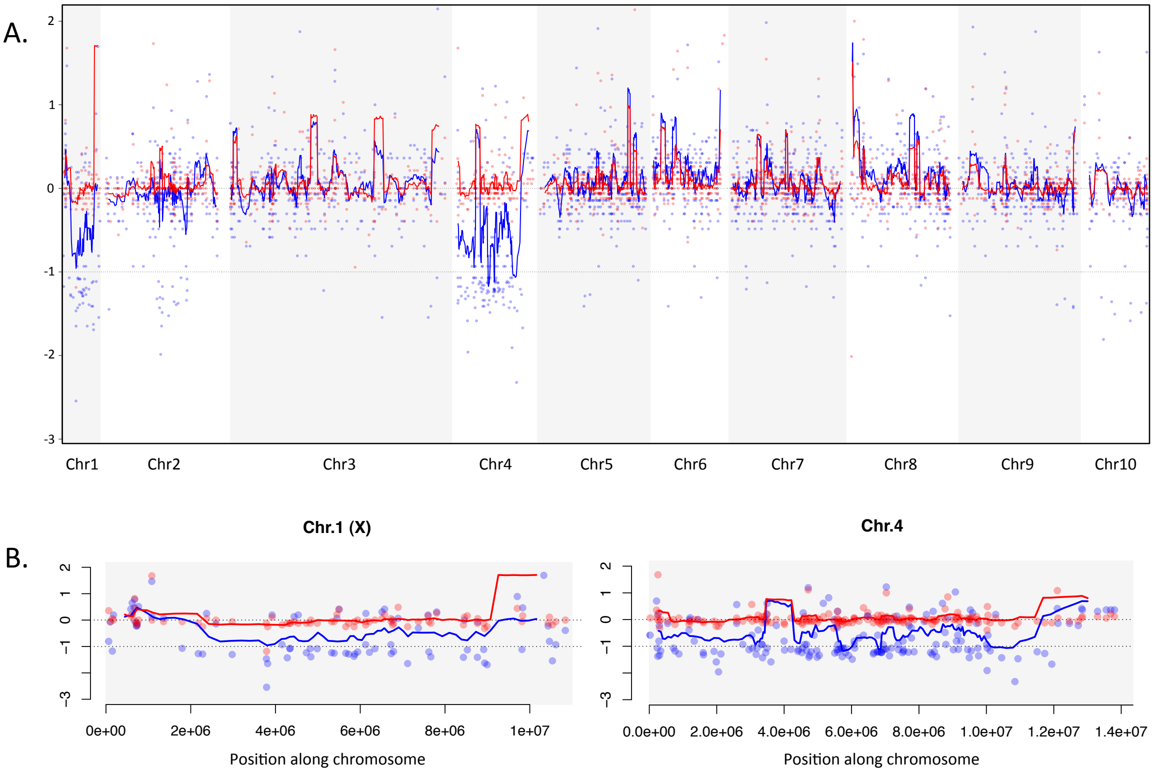


Figure 2 – Sliding window analysis of scaffolds mapped along the *Tribolium* genome. (A) Log2 of coverage densities of males (in blue) and females (in red) for the scaffolds that mapped to the ten chromosomes in *T. castaneum*. The lines represent a sliding window along the chromosomes, with a window size of 10 genes. Chromosomes 1 and 4 show a clear drop in coverage for males as compared to females. (B) Same as (A) but zoomed in into chromosome 1 and chromosome 4.

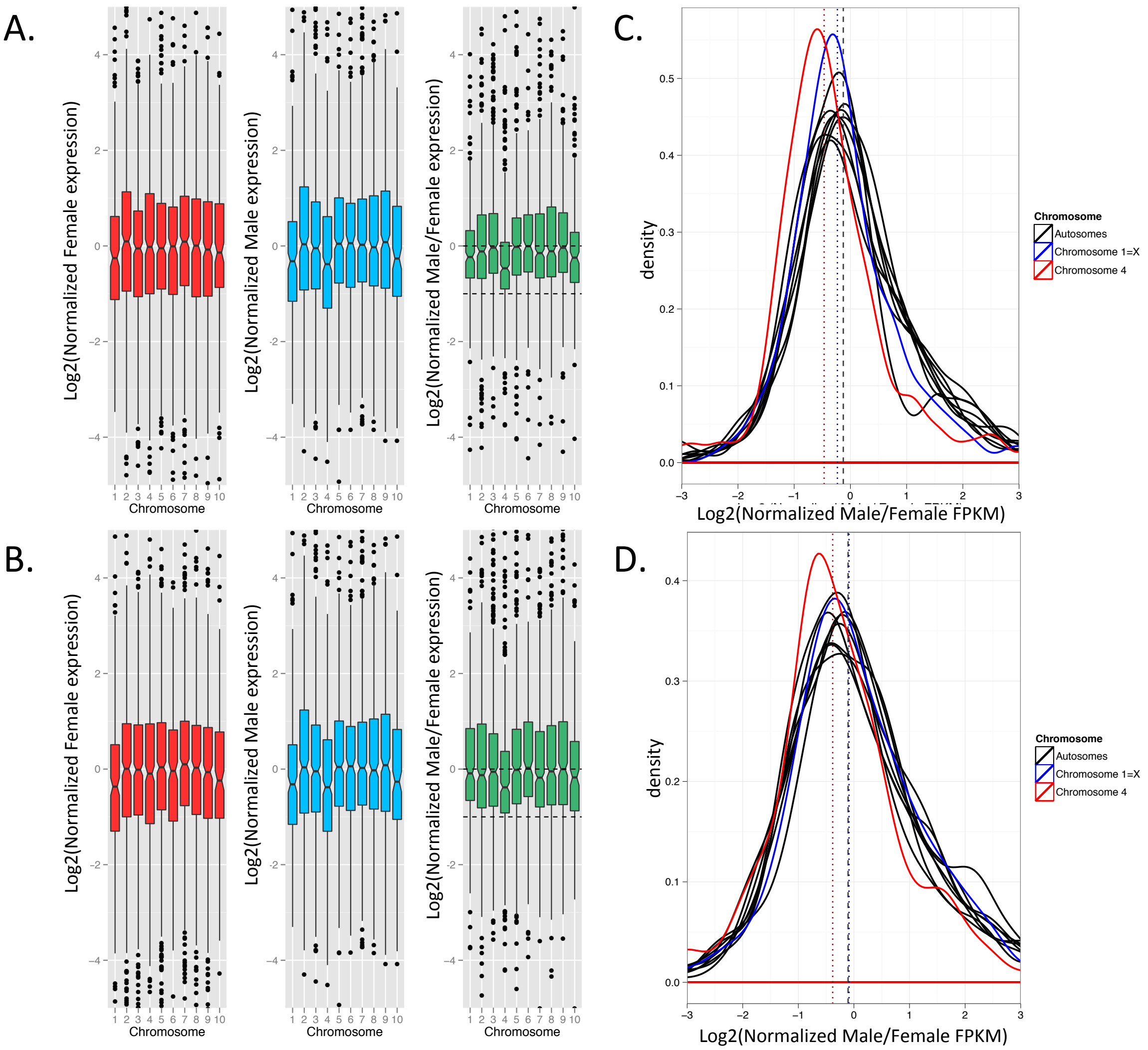


Figure 3 – Dosage compensation analysis. (A, B) Boxplot of Log2 of expression in (A) 4rd instar female larva and (B) neotenic adult female (in red), male (in blue), and male/female (in green). FPKM cutoffs for each sample were determined based on the FPKM values for the introns and intergenic regions (see **Fig. S5**). The distribution of log2(Male/Female) FPKM values for chromosome 4 is significantly different than that of the autosomes with a Wilcoxon test (p-value of < 2.2e-16 for 4th instar female larva; p-value of 4.297e-13 for neotenic adult female). There was no significant difference observed for any other chromosome in either comparison (Corrected for multiple testing). This result holds true for several FPKM cutoffs: 0, 1, 10 (see **Figs. S6, S7**). (C, D) Density plot of log2 of (C) normalized male/ 4th instar female larva FPKM and (D) normalized male/ neotenic adult female FPKM. Normalization was done by dividing each chromosome by the median of the expression of all genes in chromosomes other than chromosome 1 and chromosome 4. Chromosome 1 is almost completely dosage compensated whereas chromosome 4 is only partially compensate

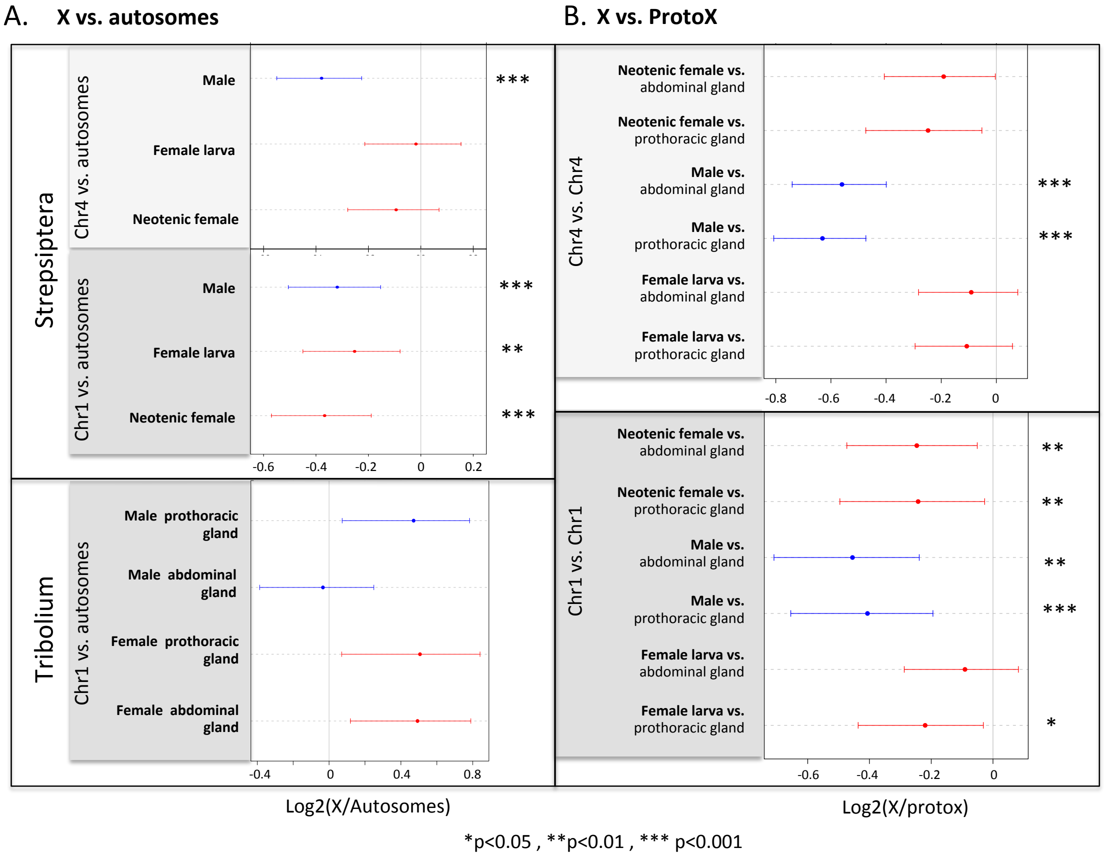


Figure 4. Current and inferred ancestral expression levels on beetle and Strepsiptera sex chromosomes. (A) X to autosome expression level ratio of expressed genes on the current sex chromosomes in *Tribolium* and *Xenos*. (B) Ancestral expression analysis of *Xenos* sex chromosomes, using expression values in male and female *T. castaneum* as a proxy for ancestral expression values. Expression of genes mapping to chromosome 4 relative to the proto X, and expression of Chr1 (which is the ancient sex chromosome) in *X. vesparum* relative to Chr1 in *T. castaneum*. Female *X. vesparum* are shown in red and male *X. vesparum* is shown in blue. Expression of each *X. vesparum* sample is compared to the expression in both abdominal (top) and prothoracic glands (bottom) of *Tribolium* for each sex separately. Asterisks are used to denote cases where a significant increase or decrease in expression is observed relative to the ancestral expression (* p-value <0.05 , ** p-value<0.01, *** p-value <0.001). Dots represent the median, and bars their approximate confidence interval (median +/- 1.57 x IQR/√n, where IQR is the interquantile range and n the sample size; this is equivalent to the notch size of a boxplot).

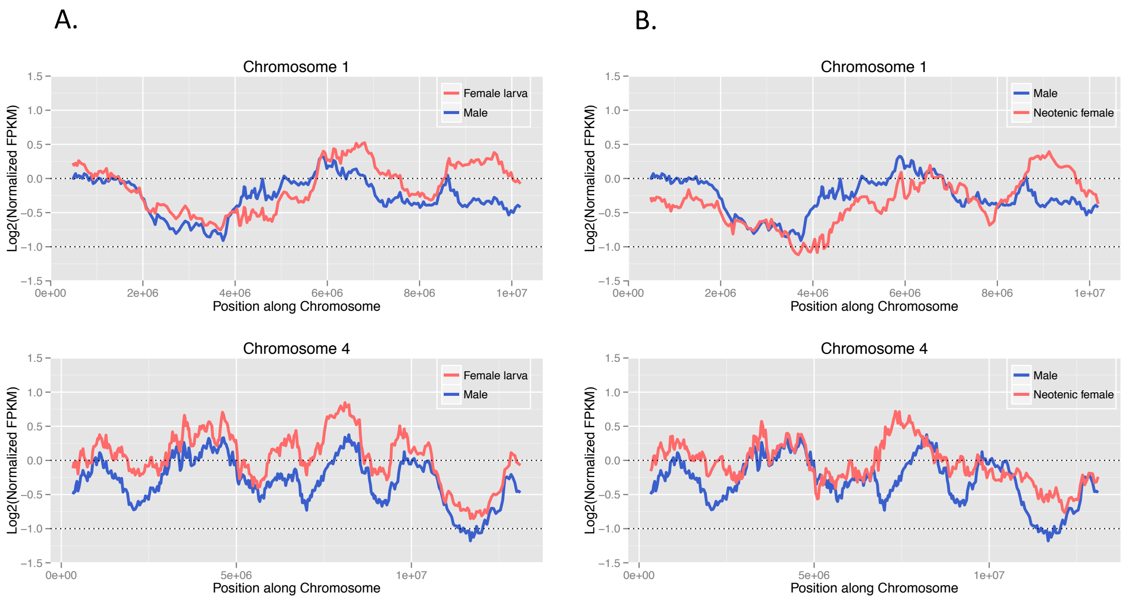
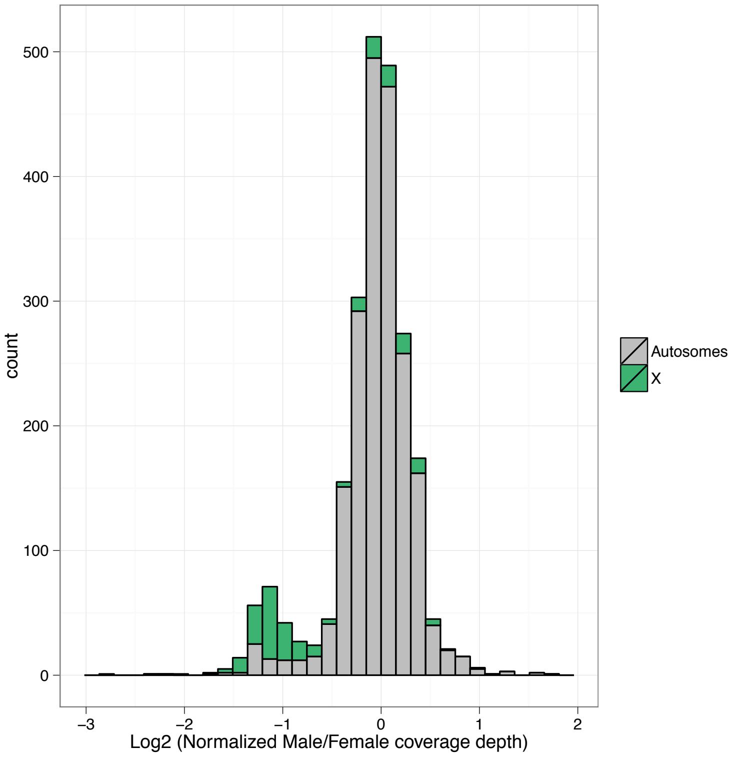
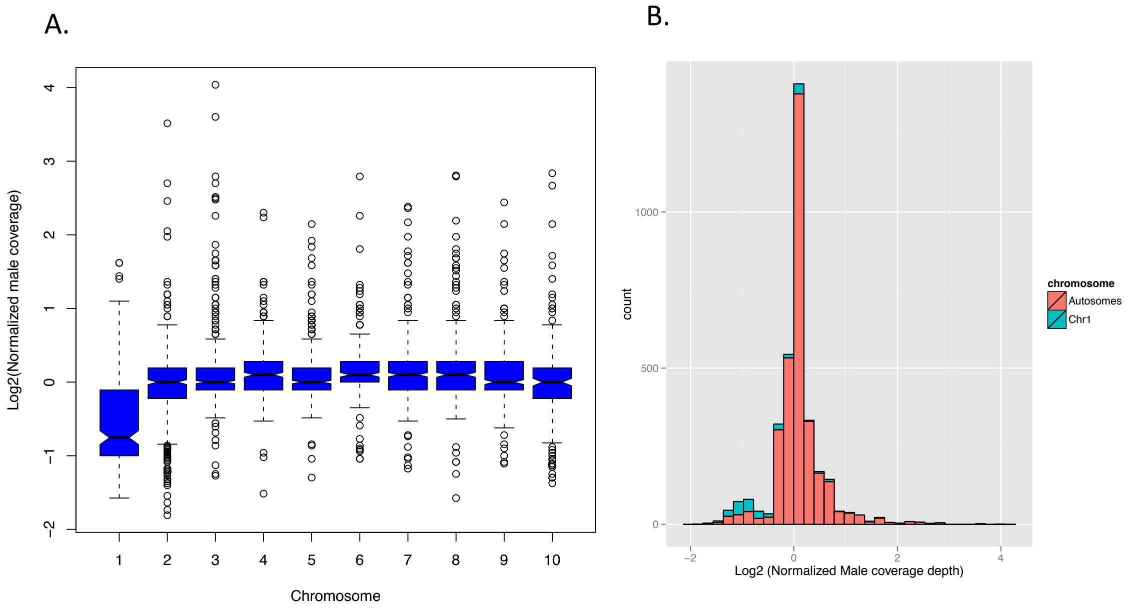


Figure 5. Sliding window analysis of dosage compensation. Log2 of normalized FPKM values of *Xenos* males and females mapped along chromosome 1 and chromosome 4 of *T. castaneum*. Male is shown in blue and (A) female larva and (B) neotenic adult female is shown in red. Chromosome 1 appears to be almost completely dosage compensated, with some regions showing male-biased and others showing female-biased expression. Chromosome 4, on the other hand, has globally lower FPKM values for males than females , suggesting that complete dosage compensation has not yet evolved on the this chromosome.

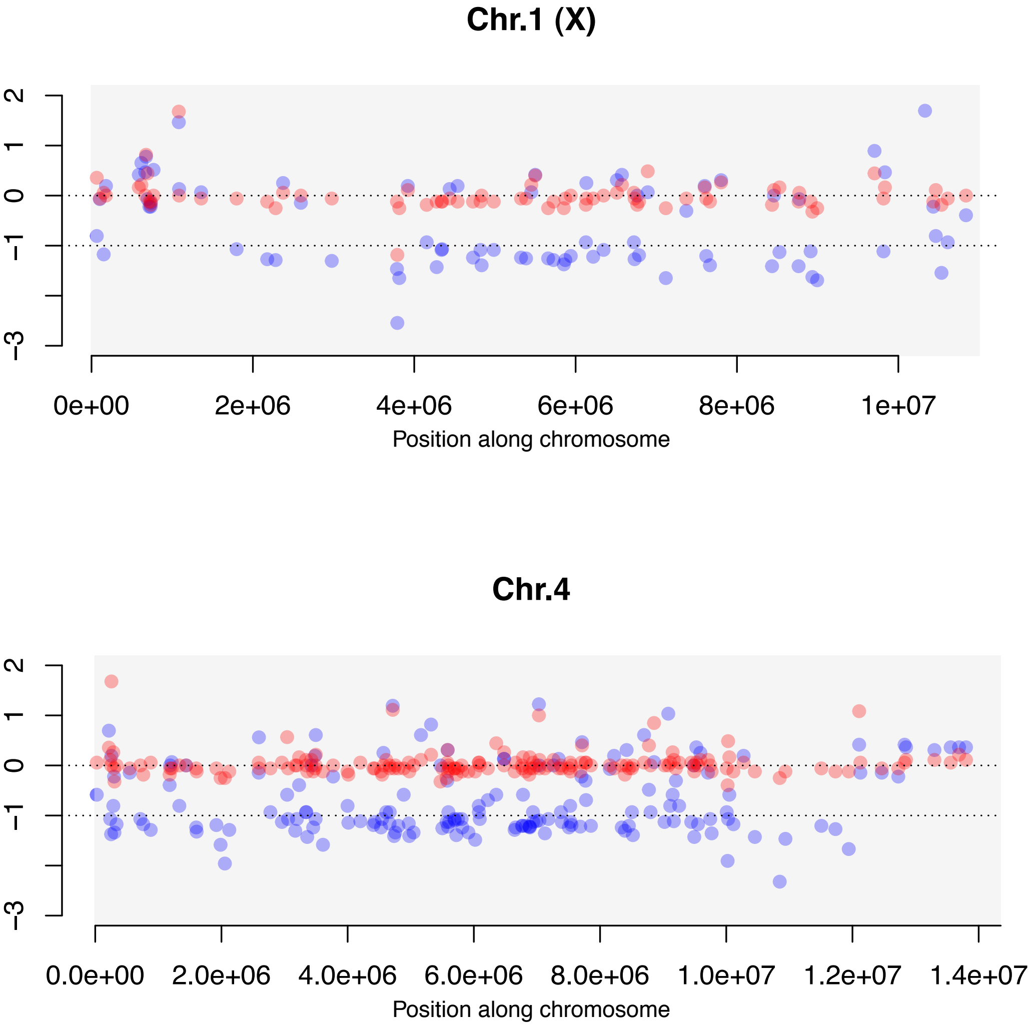
Supplementary Figures



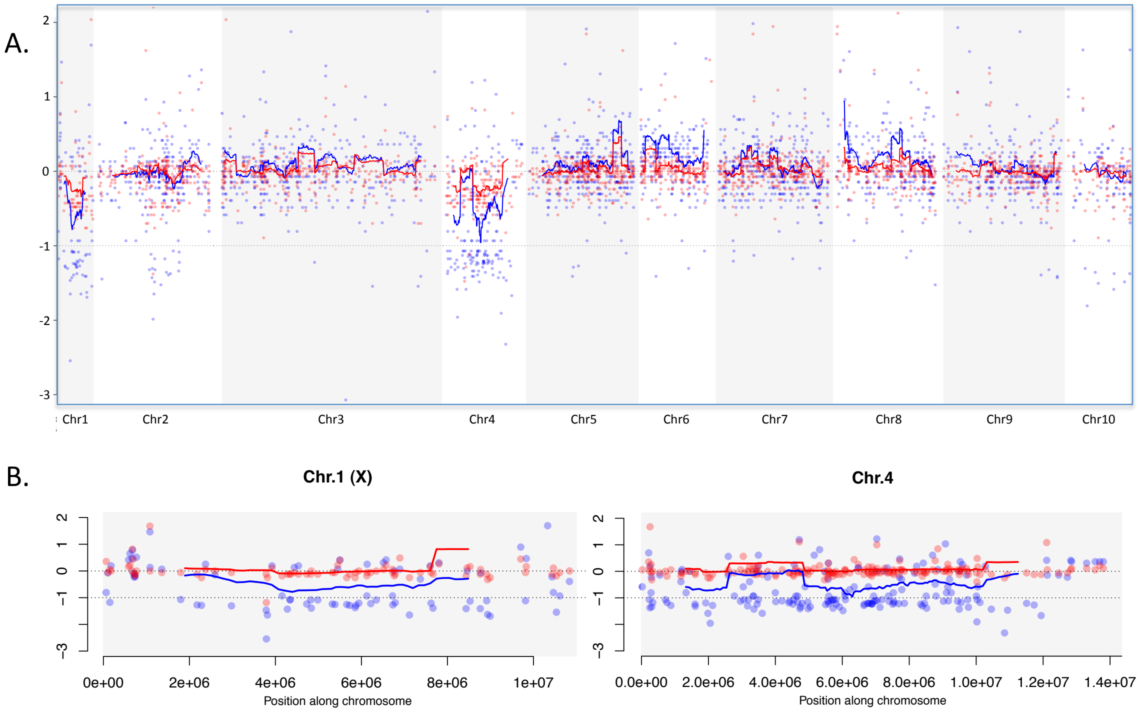
Supplementary Fig S1, Histogram of Log2 male/female coverage. Scaffolds that map to chromosomes 1 and 4 of *T. castaneum* are shown in green. The bimodal distribution in coverage suggests that a substantial fraction of the genome is sex-linked in *Xenos*, with the peak with reduced male/female coverage corresponding to scaffolds that are X-linked in *X. vesparum*.



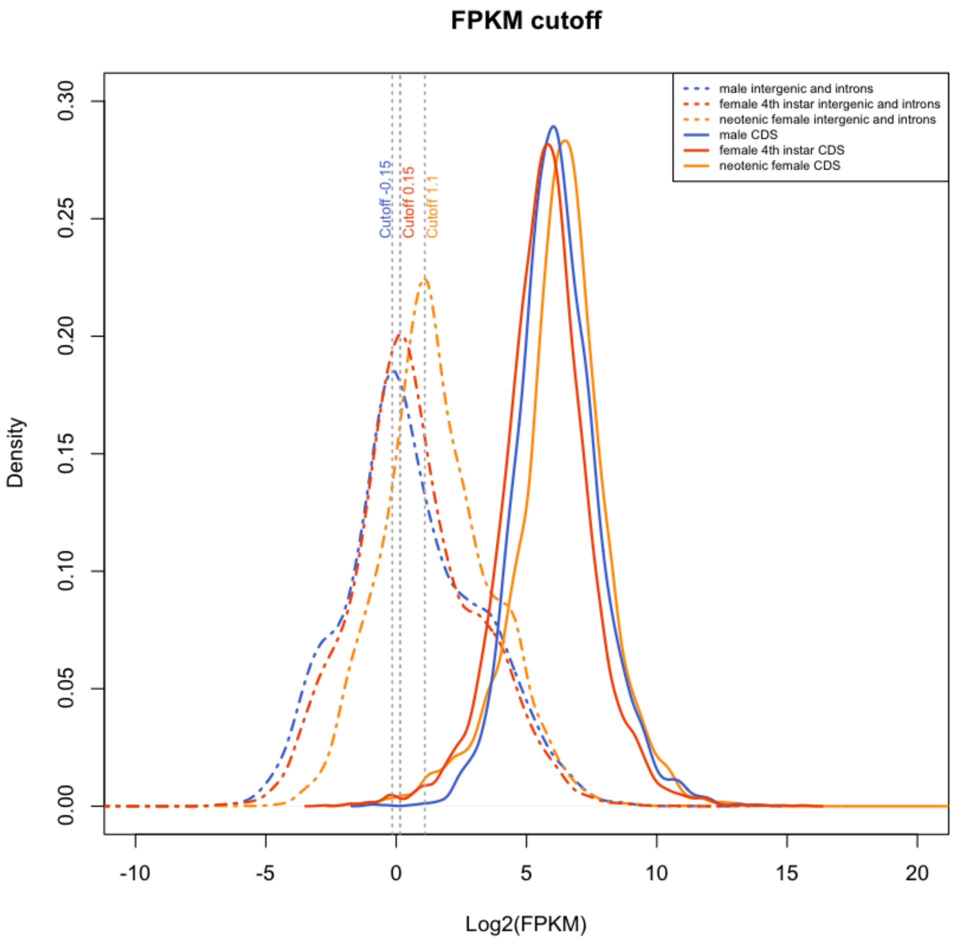
Supplementary Fig S2. Coverage analysis in *M. moldrzyki.* (A) Boxplot of Log2 of normalized male coverage for scaffolds mapping to chromosomes 1-10 of *T. castaneum.* (B) Histogram of log2 of normalized male coverage. The bimodal distribution shows clear distinction between the coverage of the X-linked scaffolds and the autosomes. Scaffolds mapping to chromosome 1 in *Tribolium* are mostly X-linked and correspond to the lower peak.



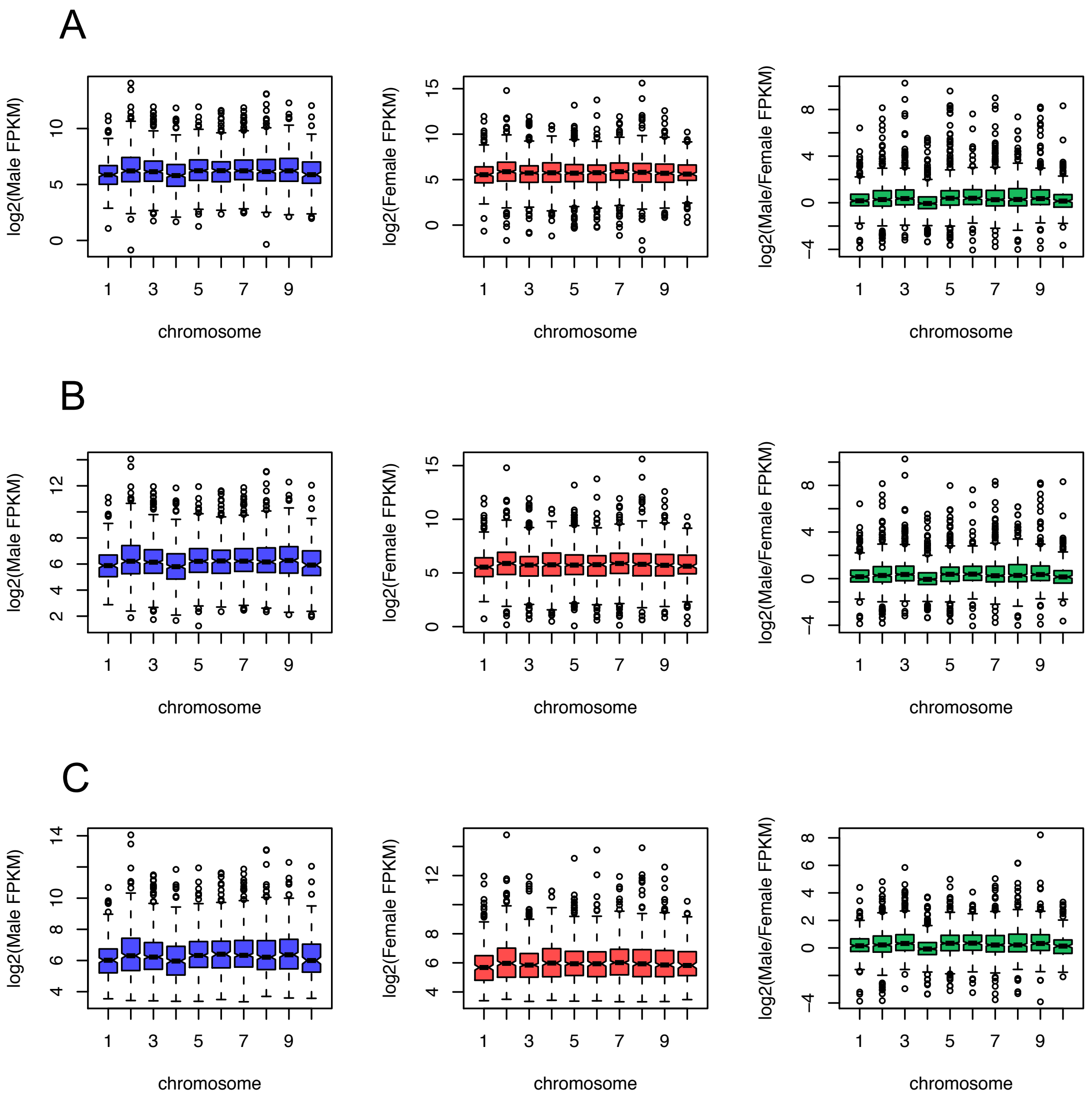
Supplementary Fig S3. Log2 of the coverage densities of males (in blue) and females (in red) for scaffolds that mapped to chromosome 1 and 4 in *Tribolium*.



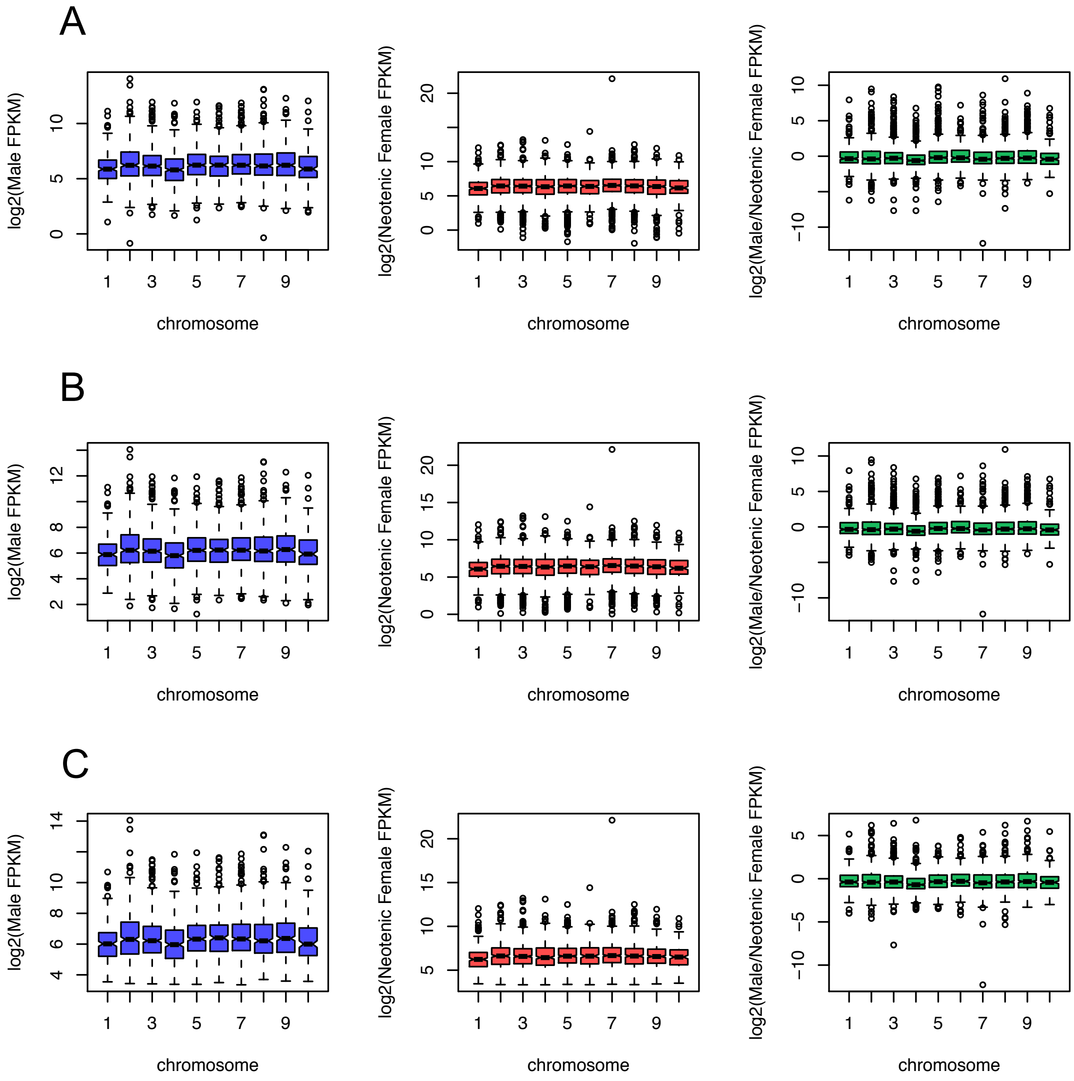
Supplementary Fig S4. Sliding window analysis of scaffolds mapped along the *Tribolium* genome. (A) Log2 of coverage densities of males (in blue) and females (in red) for the scaffolds that mapped to the ten chromosomes in *T. castaneum*. The lines represent a sliding window along the chromosomes, with a window size of 30 genes. Chromosomes 1 and 4 show a clear drop in coverage for males as compared to females. (B) Same as (A) but zoomed in into chromosome 1 and chromosome 4.



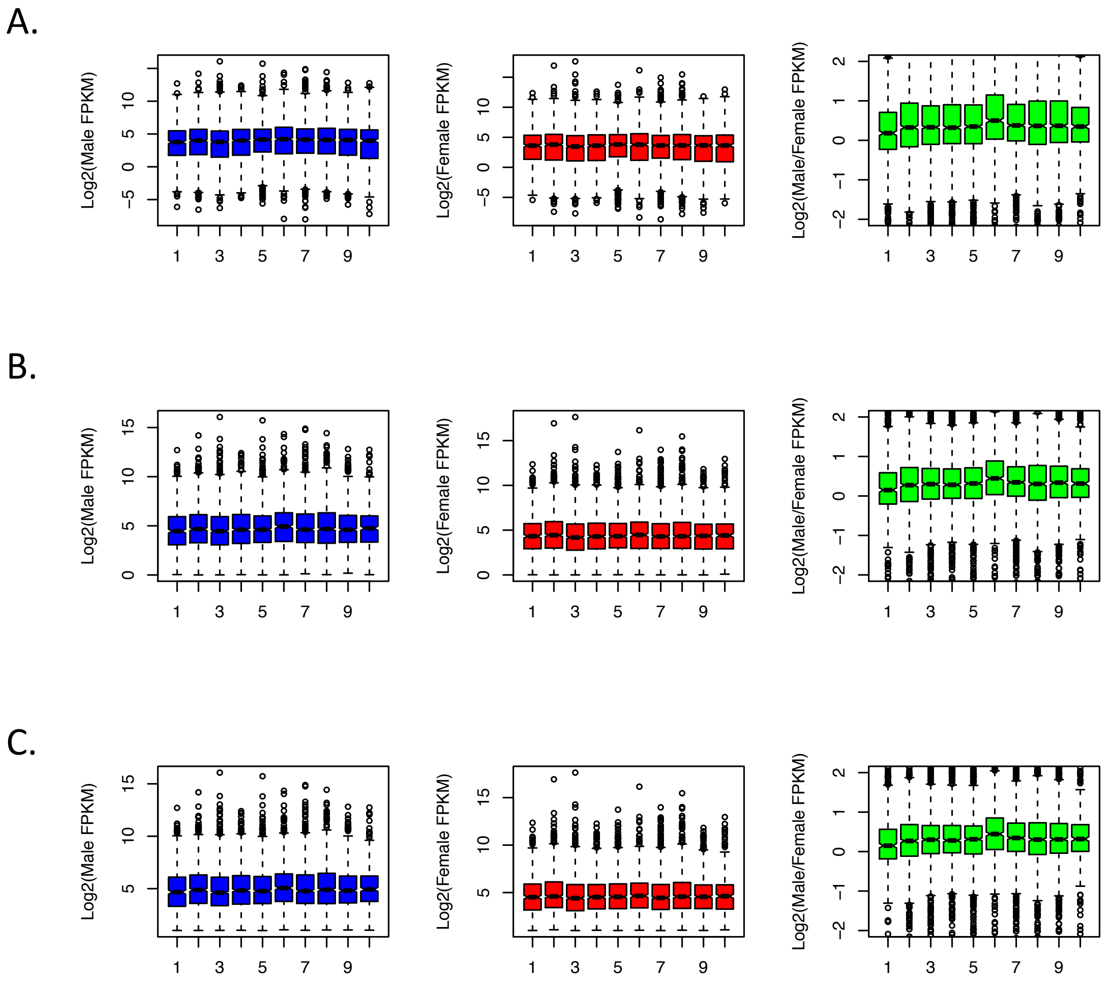
Supplementary Fig S5. Density plot of log(2) of the FPKM values for coding sequences (in continuous lines) and introns and intergenic regions (in dashed lines). Male is shown in blue, neotenic adult female is shown in orange and female larva is shown in red. The peaks for the introns and intergenic regions are used to determine the FPKM cutoff for the coding sequences for each sample.



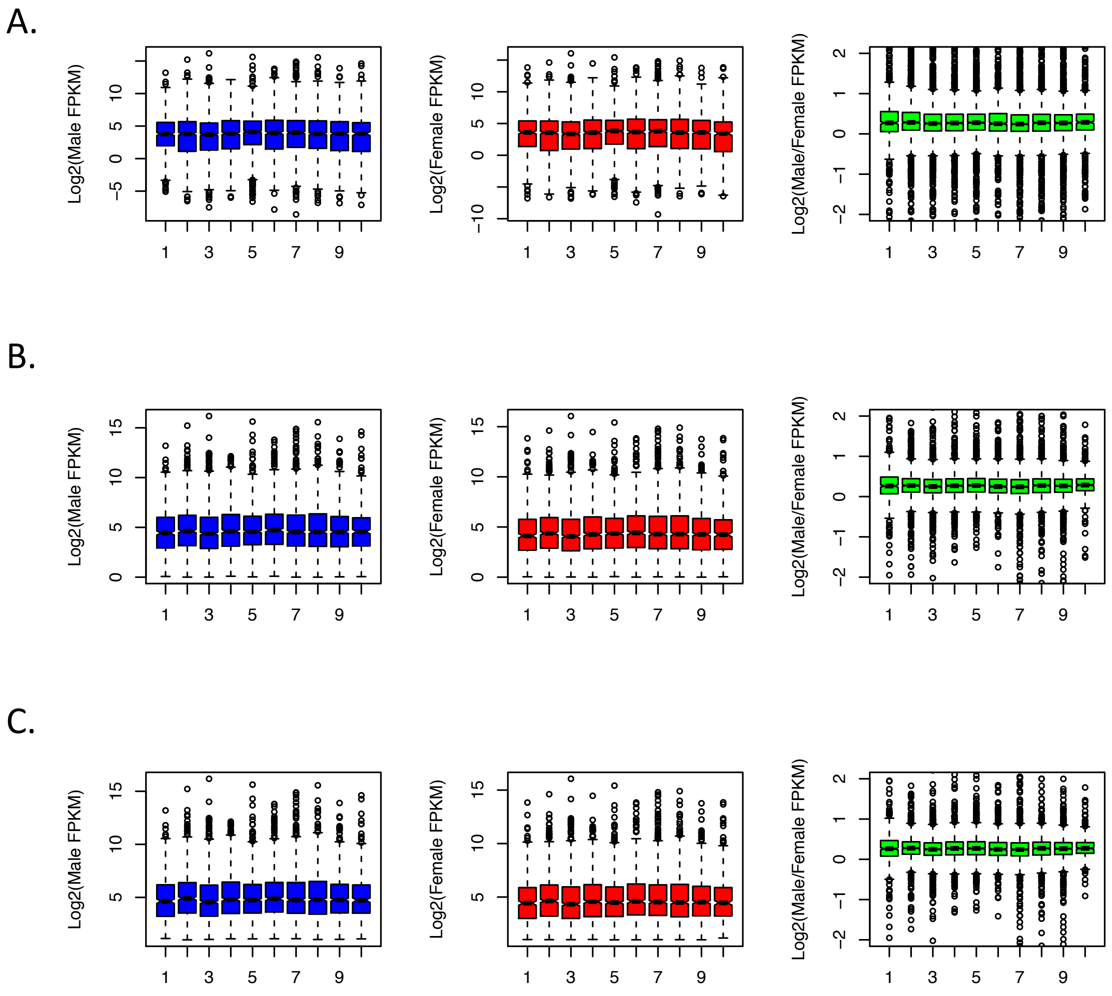
Supplementary Fig. S6. Boxplot of Log2 of the expression in the 4th instar female larva (in red), male (in blue), and male/female (in green). Three different FPKM cutoffs were used, (A) 0 (B)1 and (C) 10 . In each case for the 4th chromosome, the distribution was significantly different than the rest of the autosomes (p-values < 2.2e-16 , 2.824e-16 and < 2.2e-16 for FPKM cutoff of 0,1 and 10, respectively).



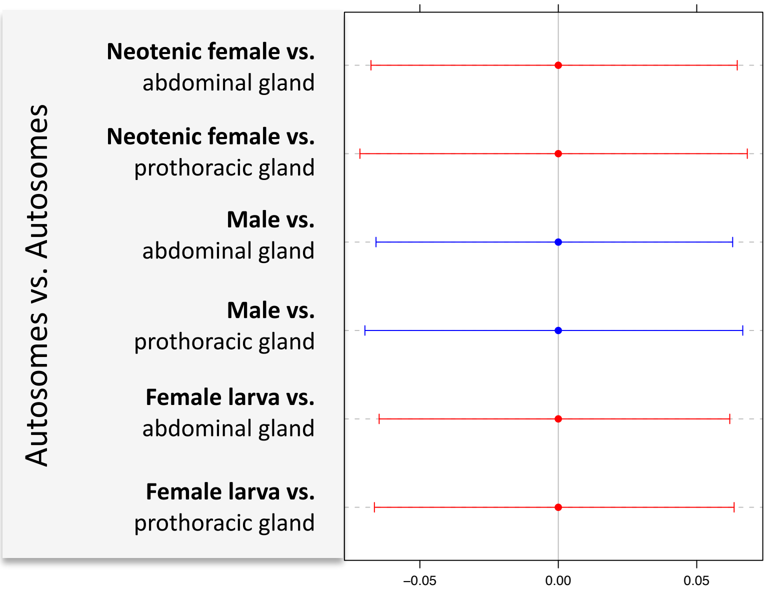
Supplementary Fig. S7. Boxplot of Log2 of the expression in neotenic adult female (in red), male (in blue), and male/female (in green). Three different FPKM cutoffs were used, (A) 0 (B) 1 and (C) 10. In each case for the 4th chromosome, the distribution was significantly different than the rest of the autosomes (p-values 1.032e-06, 1.748e-06 and 1.007e-08 for FPKM cutoff of 0, 1 and 10, respectively).



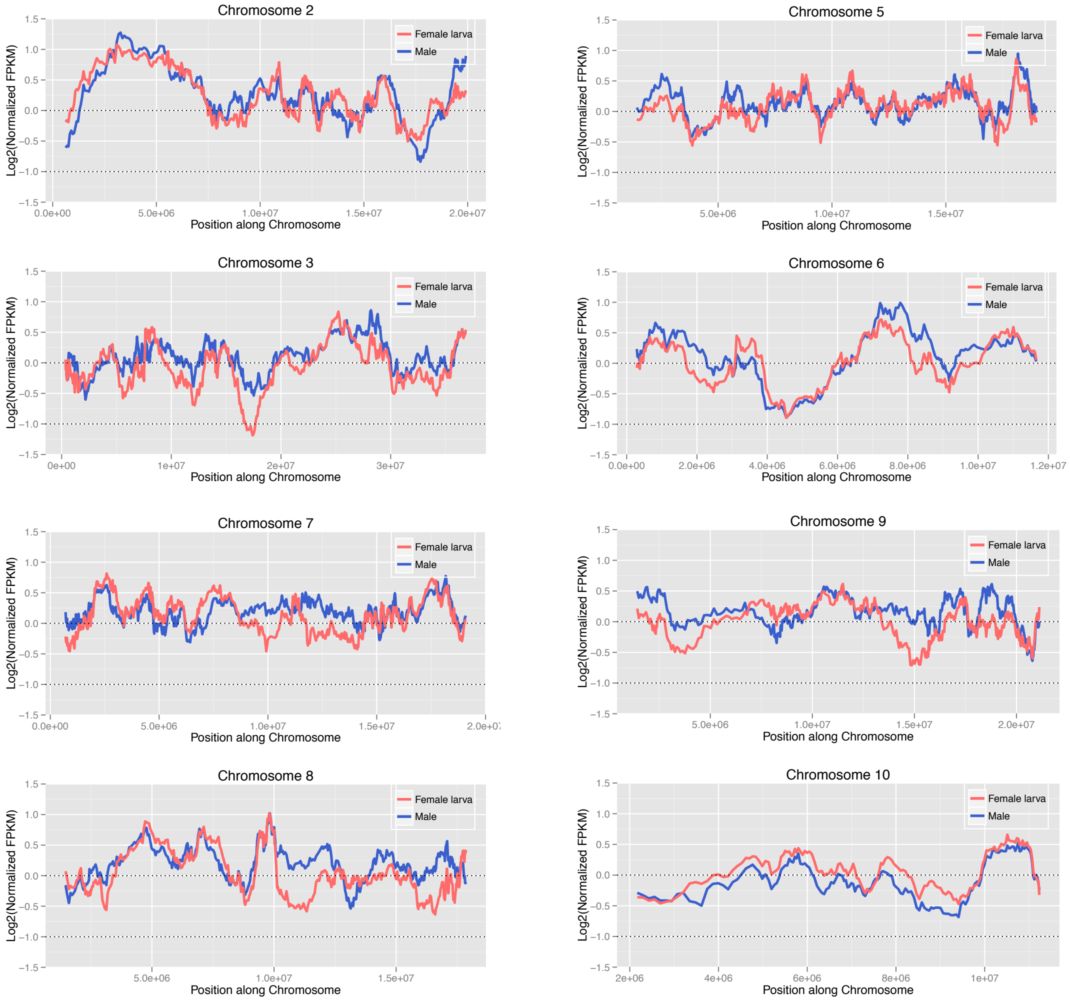
Supplementary Fig. S8. Expression analysis of abdominal glands from *T. castaneum*
Log2 of FPKM values using cutoffs (A) FPKM>0 (B) FPKM>1 and (C) FPKM>2 for FPKM for chromosome 1-10. For all three cutoffs, no significant hypertranscription of the X is detected in females and all chromosomes are expressed at similar levels. Chromosome 1 in males is expressed at a slightly lower level for all three cutoffs (Wilcoxon test p-values 0.047, 0.055 and 0.01 for FPKM cutoff 0,1 and 2 respectively when comparing the expression of chromosome 1 with that of the autosomes; and Wilcoxon test p-values 1.222e-06, 2.461e-08, 2.769e-08 for FPKM cutoff 0,1 and 2 respectively when comparing male/female expression of chromosome 1 versus the male/female expression of the autosomes.)



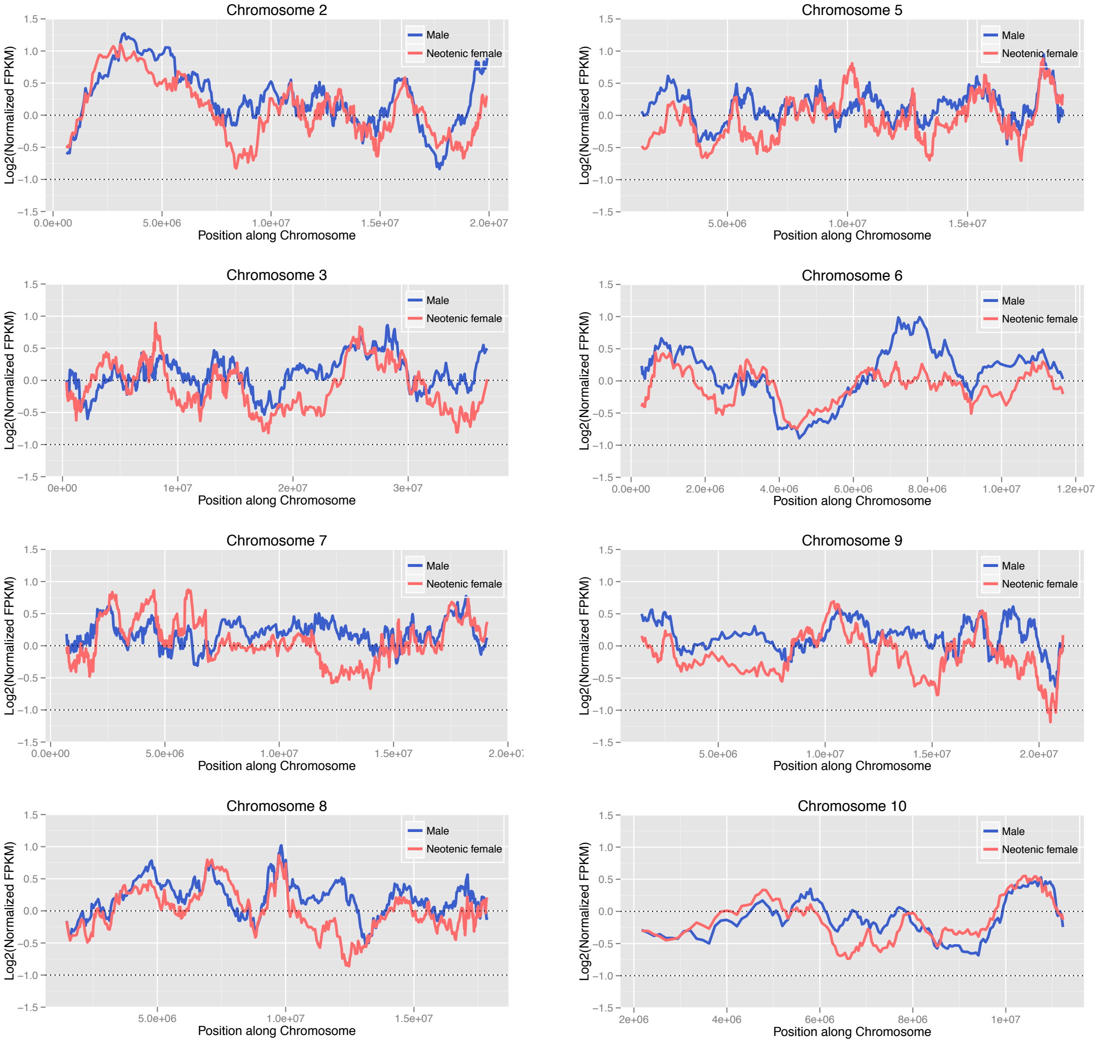
Supplementary Fig S9. Expression analysis of prothoracic glands from *T. castaneum*
Log2 of FPKM values using cutoffs (A) FPKM>0 (B) FPKM>1 and (C) FPKM>2. For both male and female, for all three FPKM cutoffs, all chromosomes are found to be expressing at similar levels. No reduction in expression is detected for chromosome 1 in males, unlike what was found for the abdominal glands (Fig. S8).



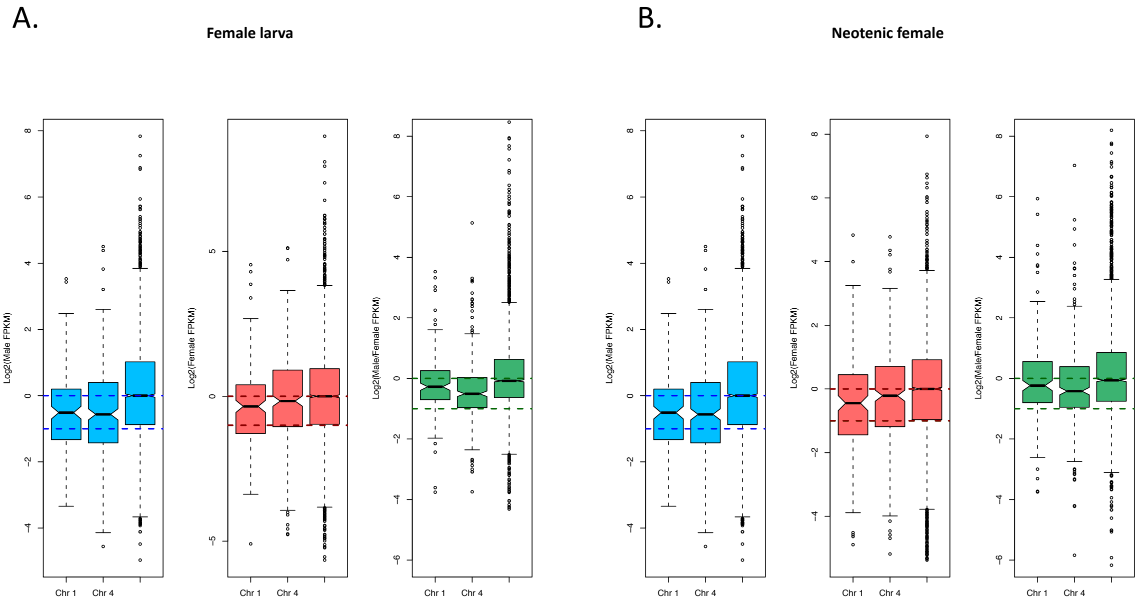
Supplementary Fig. S10. Ancestral expression analysis of Xenos sex chromosomes, using expression values in male and female *T. castaneum* as a proxy for ancestral expression values. Expression of autosomes in *X. vesparum* relative to the expression of autosomes in *T. castaneum*. Female *X. vesparum* are shown in red and male *X. vesparum* is shown in blue. Expression of each *X. vesparum* sample is compared to the expression in both abdominal (top) and prothoracic glands (bottom) of *Tribolium* for each sex separately.



Supplementary Fig. S11. Sliding window analysis of gene expression in autosomes. Log2 of normalized FPKM values of *Xenos* males and females mapped along chromosomes 2,3,5,6,7,8,9,10 of *T. castaneum*. Male is shown in blue and female larva is shown in red.



Supplementary Fig. S12. Sliding window analysis of gene expression in autosomes. Log2 of normalized FPKM values of *Xenos* males and females mapped along chromosomes 2,3,5,6,7,8,9,10 of *T. castaneum*. Male is shown in blue and neotenic adult female is shown in red.



Supplementary Fig. S13. Dosage compensation analysis for scaffolds assigned as X-linked based on coverage. (A, B) Boxplot of Log2 of expression in (A) 4th instar female larva and (B) neotenic adult female (in red), male (in blue), and male/female (in green). The distribution of log2(Male/Female) FPKM values for chromosome 4 is significantly different than that of the autosomes with a Wilcoxon test (p-value of 6.5e-16 for 4th instar female larva; p-value of 3.1e-06 for neotenic adult female).

Chapter 2

Convergent evolution of Y chromosome gene content in flies

Shivani Mahajan & Doris Bachtrog

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

Abstract

Sex-chromosomes have formed repeatedly across Diptera from ordinary autosomes, and X-chromosomes mostly conserve their ancestral genes. Y-chromosomes are characterized by abundant gene-loss and an accumulation of repetitive DNA, yet the nature of the gene repertoire of fly Y-chromosomes is largely unknown. Here, we trace gene-content evolution of Y-chromosomes across 22 Diptera species, using a subtraction pipeline that infers Y genes from male and female genome and transcriptome data. Few genes remain on old Y-chromosomes, but the number of inferred Y-genes varies substantially between species. Young Y-chromosomes still show clear evidence of their autosomal origins, but most genes on old Y-chromosomes are not simply remnants of genes originally present on the proto-sex-chromosome that escaped degeneration, but instead were recruited secondarily from autosomes. Despite almost no overlap in Y-linked gene content in different species with independently formed sex-chromosomes, we find that Y-linked genes have evolved convergent gene functions associated with testis-expression. Thus, male-specific selection appears as a dominant force shaping gene-content evolution of Y-chromosomes across fly species.

Introduction

X and Y chromosomes are involved in sex determination in many species (Bachtrog *et al.* 2014). Sex chromosomes are derived from ordinary autosomes, yet old X and Y chromosomes contain a vastly different gene repertoire. In particular, X chromosomes often closely resemble the autosome from which they were derived, with only few changes to their gene content (Vicoso and Charlesworth 2006). In contrast, Y chromosomes dramatically remodel their gene repertoire (Charlesworth and Charlesworth 2000; Bachtrog 2013; Hughes and Page 2015). 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 (Bachtrog 2013). Old Y chromosomes typically contain only a few genes, and some lineages have lost their Y chromosome entirely (Blackmon *et al.* 2017). The ultimate cause for Y degeneration is a lack of recombination on Y chromosomes, which renders natural selection inefficient (Bachtrog 2013). However, while X chromosomes have been characterized and sequenced in many species, much less is known about Y gene content evolution beyond these very general patterns. Labor intensive sequencing of Y chromosomes in a few mammal species has revealed a surprisingly dynamic history of Y chromosomes, with palindromes retarding Y degeneration in primates (Skaletsky *et al.* 2003), or meiotic conflicts driving gene acquisition on the mouse Y (Soh *et al.* 2014). However, the repeat-rich nature of Y chromosomes has hampered their evolutionary studies in most organisms.

Dipteran flies have multiple independent originations of sex chromosomes (Vicoso and Bachtrog 2015). In particular, flies typically have XY sex chromosomes and a conserved karyotype consisting of six chromosomal arms (five large rods and a small dot; termed Muller elements A-F (White 1949)). Interestingly, we recently showed that superficially similar karyotypes conceal the true extent of sex chromosome variation in Diptera: whole-genome analysis in 37 fly species belonging to 22 families identified over a dozen different sex chromosome configurations in flies based on gene content conservation of the X chromosome (Vicoso and Bachtrog 2015). The small dot chromosome was repeatedly used as a sex chromosome, but we detected species with undifferentiated sex chromosomes, others in which a different chromosome replaced the dot as a sex chromosome or in which multiple chromosomal elements became incorporated into the sex chromosomes, and others yet with female heterogamety (ZW sex chromosomes) (Vicoso and Bachtrog 2015).

However, no Y-linked genes were identified in our previous analysis, due to the difficulty in assembling genes from the often highly repeat-rich Y chromosome. Several Y-linked protein-coding genes in *Drosophila melanogaster*, for example, carry mega-base sized introns consisting of repetitive transposable element (TE) and satellite-derived DNA (Gatti and Pimpinelli 1992), making it impossible to assemble them using next-generation sequencing approaches (Carvalho *et al.* 2000; 2001) (though the application of long-read PacBio technology has proven useful in assembling Y-linked genes and genomic regions in *D. melanogaster* (Krsticevic *et al.* 2015; Carvalho *et al.* 2015)). Intriguingly, most Y-linked genes in Drosophila are not simply remnants of genes present on the autosome that became the sex chromosome; instead, they all appear to have been acquired secondarily on the Y, after it evolved its male-limited transmission (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). Y-linked genes in *D. melanogaster* all have male-specific functions and have adapted testis-specific expression, which suggests that they were acquired from autosomes and retained on the male-specific Y because of male-beneficial functions (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). This is in contrast to most mammalian species studied: while mammals have also acquired some multi-copy testis-specific genes secondarily, they still contain multiple genes that arose from genes ancestrally present on the proto-sex chromosomes with broad expression patterns and homologs on the X (Lahn and Page 1997; Skaletsky *et al.* 2003; Hughes *et al.* 2010; Cortez *et al.* 2014; Bellott *et al.* 2014). These genes may have been maintained because of dosage constraints (Cortez *et al.* 2014; Bellott *et al.* 2014).

Here, we utilize whole-genome and transcriptome sequencing data from 22 Diptera species to trace gene content evolution of Y chromosomes in flies. Our sample encompasses sex chromosomes of very different ages, and at very different stages in their evolution. Our broad phylogenetic sampling across Diptera families focuses on old, independently formed Y chromosomes that presumably have been sex-linked for long time periods (i.e. several tens of millions of years), with basically no sequence homology left between the X and the Y (Vicoso and Bachtrog 2015). Drosophila neo-sex chromosomes, on the other hand, were formed more recently (tens of thousands of years, to a few million years ago), by fusions of different autosomes to the ancestral sex chromosome pair of Drosophila (which is conserved across Drosophilidae). For recent fusions, the neo-X and neo-Y still contain considerable homology between them, and the amount of sequence homology progressively declines for older fusions as Y chromosomes degenerate (Zhou *et al.* 2012; Zhou and Bachtrog 2012; Bachtrog 2013; Zhou and Bachtrog 2015). This contrast enables us to infer the selective regime under which Y chromosomes evolve initially when still containing most of their ancestral genes, and their long-term evolutionary dynamics after most of their original genes have been lost.

In particular, our sampling scheme allows us to compare Y gene complement evolution on three different levels: (1) gene content evolution on old, non-homologous Y chromosomes across Diptera families; (2) the dynamics of gene gain and loss on the ancestral homologous Y chromosome of Drosophilidae; and (3) modification of the ancestral gene complement on young, recently formed Drosophila neo-Y chromosomes. Here, we identify Y-linked genes in 13 Diptera species, using a subtraction pipeline that infers Y genes from male and female genome and transcriptome data. We show that most Y genes in flies are derived from autosomes, and have convergently evolved males-specific functions.

Results

Inference and validation of Y-linked genes in *D. melanogaster*

Previous studies used male and female genomic data to identify Y-linked genes in *Drosophila* or *Anopheles* species (Koerich *et al.* 2008; Hall *et al.* 2013; Carvalho and Clark 2013; Hall *et al.* 2016). In particular, by comparing male and female sequence data to a reference genome, Y-linked sequences can be identified based on being present only in the male sequence data (either by identifying scaffolds with male-specific kmers (Carvalho and Clark 2013) or by finding scaffolds with higher read coverage in male relative to female genomic reads (Hall *et al.* 2013)). Our initial application of these approaches to our male and female genomic fly data was of limited success to reliably identify Y genes (Vicoso and Bachtrog 2015), presumably due to a combination of factors: Y chromosomes have few genes and mainly consist of repetitive DNA, and our genome assemblies for the various fly taxa from next-generation sequencing data are more fragmented than the well-curated *Drosophila* or *Anopheles* genomes, and especially so at repeat-rich regions. Thus, fragmented genome assemblies combined with moderate genomic read coverage prevented us from using methods to infer Y-linked genes simply based on genomic data.

Instead, we developed a bioinformatics subtraction pipeline to identify Y-linked genes, using both transcriptome and genome assemblies and raw sequencing reads from both sexes (**Fig. 1**), which is similar to an approach performed in mammals (Cortez *et al.* 2014). Briefly, male transcripts were assembled from male RNA-seq reads that did not map to a female genome assembly, and Y identity was confirmed by mapping to male genomic and transcriptomic reads, and no/little mapping to female genomic and transcriptomic reads (see **Fig. 1**, Methods).

We validated our pipeline by applying it to genomic and RNA-seq data that we collected for *D. melanogaster* males and females (**Supplementary Table 1**), and we could recover all previously identified Y genes, with the exception of the recently acquired *FDY* gene (**Supplementary Fig. 1**). *FDY* still shares considerable homology with its autosomal paralog (98% nucleotide identity (Carvalho *et al.* 2015)), and thus does not pass our strict bioinformatics filters. Our *D. melanogaster* assemblies of Y-linked transcripts are also highly contiguous and span almost all of the annotated coding sequences on the *D. melanogaster* Y chromosome (**Supplementary Fig. 1**). Most genes are covered by a single, full-length transcript, and four genes are covered by two partial transcripts with short gaps; only the *PRY* gene is missing a substantial fraction of its coding sequence in our *de novo* transcriptome assembly (the missing fragment did not pass the genomic coverage threshold in our pipeline). Moreover we were also able to recover genes from the *Mst77Y* gene family (**Supplementary Fig. 1**), which still retain moderate levels of homology to their autosomal paralog *Mst77F* (~90% identical at the protein level (Krsticevic *et al.* 2010)).

In addition to the known Y genes, we identify one previously unmapped coding transcript on the *D. melanogaster* Y that corresponds to the annotated *CG41561* gene (which was suspected to be Y-linked (Daines *et al.* 2011)). This protein-coding gene is located on an unmapped 16.1-kb long scaffold, and has four annotated coding exons (**Fig. 2A**). We confirmed Y-linkage of that gene by read mapping to other published *D. melanogaster* male and female strains: *CG41561* was present in all males sequenced from various locations, but absent in reads derived from females (**Supplementary Table 2, Fig. 2B**). This supports our conclusion that *CG41561* is Y-linked in *D. melanogaster*, and fixed among *D. melanogaster* strains. Expression profiles show that *CG41561* is expressed predominantly in testis, and to some extent also in L3 larvae (**Fig. 2C, Supplementary Fig. 2**). We could not detect a paralog in the *D. melanogaster* genome for *CG41561* (even at low stringency), and orthologs were found within the melanogaster species group of Drosophila (**Supplementary Fig. 3**). Thus, like most other *D. melanogaster* Y genes (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008), *CG41561* does not have an old X homolog, and has a male (testis) function.

To infer the false-positive rate of our approach, we applied the same subtraction pipeline to identify female-specific transcripts by switching the sexes (i.e. assemble female transcripts that map to female genomic reads, but not to a male genome assembly or genomic reads, or male transcriptome data). We identify three putative female-specific transcripts, all of which are derived from the gene *kirre* that is located on the *D. melanogaster* X chromosome, and which shows higher expression in adult females compared to males. X-linked genes have reduced read coverage in males relative to females, and are thus more likely to be mis-inferred as female-specific. Overall, our pipeline shows both high sensitivity and specificity for detecting Y-linked genes, especially for species and genomic regions with high read coverage.

Identification of Y genes across Diptera

We initially applied our pipeline to 22 fly species for which we obtained genome and transcriptome data (**Supplementary Table 1**). Inferred genome sizes vary dramatically across the species investigated (Vicoso and Bachtrog 2015) (between 103-937 Mb; **Supplementary Table 3**). Overall, the quantity and quality of data collected is roughly comparable among species and similar to the *D. melanogaster* data analyzed above (between 8–87 million genomic reads per species, with more reads collected for species with larger genome sizes; **Supplementary Table 3**), suggesting that our power and sensitivity to detect Y-linked genes in other species should be roughly similar to that in *D. melanogaster*. However, genome size and the quality of genome and transcriptome assemblies, and to some extend, genomic read coverage, differ considerably among species. For instance, N50 for genome assemblies vary between 1-242 kb (**Supplementary Table 3**), and species with larger inferred genome sizes tend to have more fragmented genomes (**Supplementary Table 3**). Thus, given the sensitivity of our pipeline to genomic coverage, and genome/transcriptome assembly qualities, we applied our method to identify both male- and female-specific transcripts for each of the species, in order to empirically assess our false-positive rate. We failed to detect male-limited transcripts in four species: *Coboldia fuscipes* (the species with the smallest and most contiguous genome); the Hessian fly *Mayetiola destructor* (where males are known to lack a Y chromosome, i.e. they are X0); *Megaselia abdita* (a species with homomorphic sex chromosomes), and the flesh fly *Sarcophaga bullata* (which has a pair of small X and Y chromosomes). In three species, we find similar numbers of male- and female-limited transcripts: *Chironomus riparius* and *Aedes aegyptii* both have homomorphic sex chromosomes (and *A. aegyptii* has the largest inferred genome size of all species analyzed; **Supplementary Table 3**); and *Condylostylus patibulatus* (a species with XY sex chromosomes, and the third largest inferred genome, **Supplementary Table 3**). We only considered species further for which we had more than twice as many male-specific than female-specific transcripts (excluding *Tipula oleracea* and *Bactrocera oleae*), thus leaving us with 13 species to identify putative Y-linked genes (see **Supplementary Table 3**).

We additionally verified that our pipeline is reliable in identifying Y-linked sequences, using two different approaches. (1) We determined the location of candidate Y-genes in a subset of species with published high-quality genomes (*Anopheles gambiae* and *Drosophila* species) and (2) we used PCR to test for male-specific amplification of candidate Y-genes for a subset of non-Drosophila flies. Consistent with the high specificity of our pipeline that we observed in *D. melanogaster*, we generally find that our candidate Y transcripts either map to previously identified Y-linked scaffolds, or to unplaced scaffolds (which likely are derived from the Y chromosome). In particular, all three candidate Y-transcripts that we found in *A. gambiae* map to the previously identified Y-linked genes *YG1* and *YG2* (Hall *et al.* 2016). Furthermore, 12 candidate Y-linked transcripts identified in *D. pseudoobscura* show highly similar sequences in the published genome (>95% of nucleotides mapping to over 50% of the transcript using blastn), and 11 of them map to unplaced scaffolds in the *D. pseudoobscura* genome. If we map putative Y-linked transcripts of its close relative *D. miranda* to the *D. pseudobscura* genome, we identify 63 transcripts that are highly similar to the reference genome sequence (>95% of nucleotides mapping over 50% of the transcript); 20 of these transcripts map to unplaced (and thus putatively Y-linked scaffolds), 39 transcripts are located on Muller element C, which is the homolog of the recently formed neo-sex chromosomes in *D. miranda* (i.e. these transcripts are presumably derived from the *D. miranda* neo-Y chromosome), and only 4 map to other genomic locations. Thus, our pipeline is highly specific in each of the species in picking up true Y-linked sequences. PCR amplification in males but not females further confirmed Y-linkage for a subset of our putative Y-linked transcripts in several non-Drosophila species (6 transcripts in *Themira minor;* 10 transcripts in *Teleopsis dalmanni;* 4 transcripts in *Ephydra hians* and 8 transcripts in *Phortica variegata*, **Supplementary Fig**. **4;** for transcripts and primers see **Supplementary Table 4**). To empirically assess a 'worst-case scenario' false-positive and false-negative rate, we subsampled our *D. melanogaster* data to match read counts with the species for which we have the lowest number of read pairs (*Mayetiola destructor*). Using this reduced dataset, we identify 25 male-specific and zero female-specific transcripts with our pipeline (we lose the *PRY* gene completely and fragments of other a few other transcripts; see **Supplementary Fig. 5**). Thus, this further suggests that our approach is robust and sensitive to infer Y-linked transcripts across the species investigated.

Fly gene repertoires

In our study, we consider Y chromosomes at two very different stages of their evolution: old ancestral Y chromosomes from diverse Diptera families where most original Y genes have been lost (Vicoso and Bachtrog 2015), and young neo-sex chromosomes of Drosophila, which may still contain most of their original genes (Carvalho and Clark 2005; Zhou *et al.* 2012; Zhou and Bachtrog 2012; 2015).

In total, we identified 187 protein-coding transcripts (or parts of transcripts), and 656 non-coding transcripts across all species that are potentially Y-linked. Note that the method that we use for classifying transcripts into coding vs. noncoding (i.e. Coding Potential Calculator (Kong *et al.* 2007)) is conservative in evaluating coding capacity of a DNA sequence, resulting in the assignment of a large number of transcripts as non-coding.

Fragmented protein-coding transcripts, or short and highly divergent proteins (as is the case for many testis-expressed transcripts, see below) may be annotated as non-coding, and Coding Potential Calculator indeed called some incomplete Y-linked transcripts of *D. melanogaster* as non-coding, even though they mapped to parts of known protein-coding Y genes. The number of inferred Y genes varies substantially between species, with no protein-coding genes identified in *Clogmia albipunctata*, and 59 potentially protein-coding transcripts found in *D. miranda* (**Fig. 3**). We identify both Y-linked genes in species without morphologically distinguishable sex chromosomes, such as in black or sand flies, but also fail to detect Y genes in others with differentiated X and Y sex chromosomes (and high-quality genomes), such as in *Coboldia fuscipes* (**Supplementary Table 3**).

We previously showed that sex chromosomes originated independently in several fly families, and different chromosomes (termed Muller elements A-F; see **Fig. 3**) are segregating as the X chromosome in different species investigated (Vicoso and Bachtrog 2015). Apart from the ancestral Y of Drosophilidae, all other sex chromosome systems investigated here evolved independently (Vicoso and Bachtrog 2015). Drosophilidae are classified into two subfamilies, Drosophilinae and Steganinae, and we showed that the X chromosome of the two subfamilies is homologous (i.e. derived from Muller element A (Vicoso and Bachtrog 2015)). Two genes that are Y-linked in *D. melanogaster*, *CCY* and *kl-2*, are also found in our list of putative Y-linked genes in *Phortica variegata* – a species from the Steganinae subfamily, indicating that they have been acquired on the Y chromosome in a common ancestor of both subfamilies. We also identified several Y-linked genes of *D. melanogaster* on the homologous Y chromosomes of *D. albomicans* (*kl-2* and *kl-3*) and *D. busckii* (*kl-2*, *kl-3*, *kl-5*, *ORY* and *PPr-Y*). As previously shown, the Y chromosome of flies from the *D. pseudoobscura* group is not homologous to the Y of *D. melanogaster* (Carvalho and Clark 2005), and none of the ancestral Y genes in Drosophila are found among our putatively Y-linked genes in *D. pseudoobscura* and *D. miranda* (**Fig. 3**). Since sex chromosomes evolved independently in the other families of flies, we expect the gene content to differ among independently evolved Y chromosomes. Indeed, putative Y-linked genes identified in non-Drosophila species show no overlap; the only exception is *CCY*, which is found on both the Y chromosome of Drosophila (where Muller element A formed the sex chromosome pair), and also on the Y chromosome of the stalked-eyed fly *T. dalmanni* (where Muller element B formed the sex chromosome pair). Thus, this suggests that *CCY* was gained independently on both the Y chromosome of Drosophilidae, and the Y of stalk-eyed flies.

Y chromosomes may contain master sex determination genes, and in some cases, we could identify potentially interesting candidate genes for further study. In *Chaoborus trivittatus,* a species with homomorphic sex chromosomes, we were able to identify 17 potentially Y-linked transcripts, one of which is homologous to the DSX protein of several other Diptera species. The *dsx* gene is involved in sex determination in flies, and *dsx* homologs are expressed in the developing gonad of many animals, and have been utilized as master sex determination genes in both vertebrates and invertebrates (Bachtrog *et al.* 2014). We also recovered the *YG2* gene in *A. gambiae*, which is thought to be the male-determining gene in this species (Hall *et al.* 2016; Krzywinska *et al.* 2016).

Origin of fly Y genes on ancestral Y chromosomes

Y-linkage of genes could be a consequence of them being ancestrally located on the autosome that became a sex chromosome and escaping degeneration, or because genes were recruited to the Y chromosome secondarily (by translocations or transpositions) only after it became male-limited (as appears to be the case for most Y-linked genes in *D. melanogaster* (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). If current Y genes represent escapees of genes ancestrally located on the sex chromosomes, we expect that their closest paralogs in the genome map to the X. In contrast, if they secondarily moved onto the Y chromosome, we expect their closest paralogs to be autosomal. Note that we cannot distinguish genes that have been copied and moved to the Y from the X secondarily from those that were ancestrally located on the Y chromosome, based on location information alone (that is, we may overestimate the number of genes being ancestrally Y-linked).

We assessed the origin of our putative Y-linked genes in Diptera using two completely independent approaches. (1) We determined on which Muller element the closest homologs of Y-linked candidate genes in *D. melanogaster* are located. (2) We investigated whether the closest paralogs of putative Y genes within the same genome are X-linked or autosomal, based on genomic coverage analysis (Vicoso and Bachtrog 2015). If current Y-genes are remnants of genes ancestrally present on the proto-Y chromosome, we expect them to map to the same Muller element (s) that formed the X chromosome in a species (using mapping information from *D. melanogaster*), and their closest paralog in the genome should be located on a genomic scaffold with half the male/female coverage ratio relative to autosomal ones (i.e. X-linked (Vicoso and Bachtrog 2015)). In contrast, if Y-genes were acquired from autosomes, we expect them to map to different Muller elements in *D. melanogaster* than the one (s) that formed the X chromosome, and their closest paralogs within a genome should harbor male/female genomic coverage ratios typical of autosomes. Note that mapping to *D. melanogaster* (i.e. our first approach) assumes conservation in gene content of Muller elements across Drosophila, which has been found to largely hold true using comparative mapping (White 1949) and whole-genome re-sequencing studies (Holt *et al.* 2002), and was validated by our previous comparative study inferring sex chromosomes across Diptera (where the vast majority of genes inferred as X-linked in various Diptera species, based on genomic coverage, mapped to a particular Muller element in *D. melanogaster*, and only few genes from other Muller elements, based on homology to *D. melanogaster*, were inferred to be X-linked based on coverage analysis (Vicoso and Bachtrog 2015).

The suborder Nematocera is distantly related to fruit flies, and we detect only one homolog of a putative Y-linked gene in Drosophila (for a species with homomorphic sex chromosomes; see **Fig. 3**). We identify paralogs within the genome for three Nematocera species (two with homomorphic sex chromosomes, and one with heteromorphic sex chromosomes). The only paralog that we identify in a species with heteromorphic sex chromosomes (i.e. a transcript that partially overlaps with the *YG1*/*YG2* genes in *A. gambiae*) is located on a scaffold with male/female genomic coverage ratios typical of the X (**Fig. 4**), and mapping of this Y-linked transcript against the *A. gambiae* genome (https://www.vectorbase.org/) also confirms that its closest non-Y-linked paralog is located on the X chromosome of *A. gambiae* (**Supplementary Fig. 6;** but note that the longer *YG1* and *YG2* transcripts mapped to autosomal locations in *A. gambiae* (Hall *et al.* 2016)). Indeed, a recent study utilizing a comprehensive RNA-seq dataset of sexed *A. gambiae* across development and whole and dissected adults (52 datasets in total) identified 8 putative Y-linked genes (including the *YG1* and *YG2* genes), and found them all to be derived from autosomes (Hall *et al.* 2016). It will be of interest to study additional Nematocera species with heteromorphic sex chromosomes, to better understand gene content evolution of the Y in this suborder.

Across most species belonging to the suborder Brachycera with old Y chromosomes (i.e. excluding *Drosophila* neo-sex chromosome systems), we find that putative Y-linked genes often have their homologs in *D. melanogaster* on several different Muller element's, and there is no overall enrichment for Y genes being derived from the same Muller element (s) that formed the X chromosome within a species (**Fig. 3**). Also, Y-linked genes have their closest paralog within a genome generally map to scaffolds that have male/female genomic coverage ratios typical of autosomes (**Fig. 4**). The Y chromosome of scavenger flies (*T. minor*), however, shows a somewhat different pattern: here, half of the identified putative Y-linked transcripts have their closest homolog map to the same Muller element that formed the X chromosome (3 out 6; **Fig. 3**), and 3 out of 4 paralogs of Y-linked transcripts show male/female coverage ratios in *T. minor* that are typical of X chromosomes (**Fig. 4**). Thus, a large fraction of Y-linked transcripts in scavenger flies may be remnants of genes initially present on the Y, while most putative Y-linked genes of stalk-eyed flies, shore flies, and Drosophilidae are derived from autosomes (consistent with *D. melanogaster* data (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008); see **Figs. 3 & 4**). Hence, unlike in mammals, ancestral Y genes in flies are often derived from a wide variety of autosomal genes that were acquired on the Y chromosome only after it became male-limited.

Sequence divergence between putative Y genes and their autosomal paralogs allow us to roughly date when genes were acquired on the Y chromosome, with more recent acquisitions showing higher amounts of sequence similarity (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). We determined protein-coding paralogs for each putative Y-linked transcript in the female genome, and calculated rates of synonymous and non-synonymous substitutions between the Y-linked transcripts and their closest paralog in the female genome assembly (**Fig. 5**). In general, for non-Drosophila species, divergence between Y-linked genes and their autosomal paralogs is relatively low (Ka from 0.038 to 0.773 and Ks from 0.039 to 4.022), compared to divergence levels inferred in *D.*

melanogaster (median Ka=0.353, Ks=4.112). Since we use *D. melanogaster* proteins to scaffold transcripts, the transcriptome assemblies for Drosophila species are more contiguous compared to the other species, which might make it more difficult to pick up more diverged paralogs in non-Drosophila flies. In general, we see a broad spread of divergence values between Y-linked genes and their paralogs, suggesting that genes were acquired on the Y chromosome at different evolutionary time points. This is consistent with patterns of gradual gene acquisition found on the Drosophila Y chromosome (Koerich *et al.* 2008).

Gene content evolution of Drosophila neo-sex chromosomes

All *Drosophila* species investigated, apart from *D. melanogaster*, harbor neo-sex chromosomes. Here, fusions between the ancestral sex chromosome of *Drosophila* and an autosome incorporated a new chromosomal arm into the ancestral sex chromosome, at different evolutionary time points. The neo-sex chromosomes of the species we investigated form a temporal gradient and display various levels of degeneration. Unlike the ancestral Y chromosome of *Drosophila*, the gene repertoire of young neo-Y chromosomes still reflects their ancestral gene complement (Zhou *et al.* 2012; Zhou and Bachtrog 2012; 2015). Our transcriptome analysis identifies some of the neo-Y genes as Y-linked transcripts, suggesting that they are sufficiently diverged at the DNA sequence level from their neo-X homologs to be identified by our bioinformatics pipeline. Indeed, for the species where a gene-rich autosome (i.e. not Muller element F) formed the neo-sex chromosomes, we generally see an overrepresentation of Y-linked genes derived from that Muller element that fused to the ancestral sex chromosome (**Fig. 3**). This suggests that they are remnants of genes originally present on the neo-Y.

The *D. albomicans* neo-X and neo-Y were only formed about 100,000 years ago, by the fusion of a large autosome consisting of Muller elements C and D to the ancestral sex chromosome, causing roughly 5000 genes to become sex-linked (Zhou *et al.* 2012). The neo-sex chromosomes of *D. albomicans* are still mostly homologous, with very little differentiation and degeneration of its neo-Y (Zhou *et al.* 2012). Our bioinformatics pipeline identified 61 Y-linked transcripts in *D. albomicans*. For 35 putative Y-linked transcripts for which we could identify paralogs in the female genome, 34 were added by the neo-Y fusion and one transcript is homologous to *kl-2* (we could not identify a paralog for the ancestrally Y-linked *kl-3* gene in the female genome of *D. albomicans*). Sequence divergence between putative neo-Y genes and their neo-X homologs is much lower (median Ka= 0.04 and median Ks= 0.21) than divergence between ancestral Drosophila Y genes and their paralogs (**Fig. 5**), consistent with the recent formation of the neo-sex chromosomes in *D. miranda*.

Drosophila busckii's neo-sex chromosome system was formed by the fusion of the small dot chromosome (Muller element F, which contains only about 100 genes) to the ancestral sex chromosomes about 1MY ago, and it displays intermediate levels of Y degeneration (Zhou and Bachtrog 2015). Detailed molecular analysis suggested that the majority of neo-Y linked genes are still present, but about half appear pseudogenized (Zhou and Bachtrog

2015). Our bioinformatics pipeline identified 139 putatively Y-linked transcripts in *D. busckii*, and for 48 of those transcripts were we able to identify paralogous sequences in the female genome; two were added by the neo-Y fusion, 16 were ancestrally Y-linked, 21 autosomal, 2 from the ancestral X and 7 whose genomic location could not be determined based on mapping to their published genome (Zhou and Bachtrog 2015), or homology with *D. melanogaster* coding sequences.

Drosophila pseudoobscura harbors an older neo-sex chromosome which arose about 15 MY ago (and which it shares with *D. miranda*). This system arose by the fusion of Muller element D (which contains roughly 3000 genes) to the ancestral X chromosome, and the fused arm is referred to as chromosome XR in the *pseudoobscura* group. Genes located on chromosome XR all appear hemizygous (Zhou and Bachtrog 2012), and the evolutionary fate of the neo-Y of the *D. pseudoobscura* group has been unclear. Intriguingly, it has been shown that the ancestral Y of Drosophila became linked to an autosome in an ancestor of the *D. pseudoobscura* species group, at around the same time when the Muller element D – X chromosome fusion occurred (Carvalho and Clark 2005; Larracuente *et al.* 2010). Consistent with this scenario, we do not detect any ancestral *Drosophila* Y genes as sex-linked in either *D. pseudoobscura* or *D. miranda* (see **Fig. 3**). Since flies in the *pseudoobscura* group contain a morphologically distinguishable Y chromosome, it had been speculated that the current Y is the unfused neo-Y, i.e. the degenerated remnant of Muller element D (Carvalho and Clark 2005). Proof for this hypothesis, however, is lacking. Indeed, we find that many (13 out of 30) of the putative Y-linked transcripts that have mapped paralogs in the *D. pseudoobscura* genome are derived from chromosome XR (i.e. Muller element D). In addition, we identify 11 Y-linked genes in *D. pseudoobscura* that have homologs in *D. melanogaster*, and seven of them are located on Muller element D. This supports the idea that the current Y of *D. pseudoobscura* is derived from the unfused neo-Y. Interestingly, three of the seven Y genes that were ancestrally present on Muller D (i.e. also linked to Muller element D in *D. melanogaster*) have lost their former homologs on chromosome XR in *D. pseudoobscura*. Several studies have shown that X chromosomes in Drosophila are an unpreferred location for genes with male-specific function (Sturgill *et al.* 2007; Assis *et al.* 2012), and all three genes that have been lost from XR are expressed predominantly in testis (both in *D. melanogaster* and *D. pseudoobscura*). Thus, 'demasculinization' of the X chromosome will further contribute to erode any remaining homology between the X and the Y.

Drosophila miranda contains two neo-sex chromosomes that originated through independent fusions at different time points. It shares the ancient neo-X fusion with *D. pseudoobscura* (i.e. chromosome XR*)*, and 25 different transcripts (corresponding to 6 genes) of the Y-linked transcripts in *D. pseudoobscura* are also Y-linked in *D. miranda* (20 of which are from Muller element D). Furthermore, *D. miranda* also harbors a more recently formed neo-sex chromosome: Muller element C became part of the ancestral Y chromosome only about 1.5 MY ago and has undergone massive degeneration, with over half of its genes pseudogenized, and ~150 genes (of the roughly 3000 genes initially present on the more recently added neo-Y) have become deleted (Zhou and Bachtrog 2012). Consistent with its intermediate level of differentiation, we identify the largest number of Y-linked transcripts in *D. miranda*: there are still many genes left on the neo-Y, and neo-Y genes are diverged enough from their neo-X homologs to be detectable by our bioinformatics approach. We identified 122 transcripts with homologous sequences in the female genome, 10 of which are located on chromosome XR (and thus are supposedly from the 'ancestral' neo-Y fusion), 21 transcripts have been acquired from autosomes/the ancestral X of *Drosophila*, and 91 transcripts whose closest paralog is located on the neo-X. Also, the majority of genes with homologs in *D. melanogaster* map to Muller element C (18 out of 29). Again, sequence divergence for the young neo-Y genes (median Ka=0.069 and Ks=0.094) is lower than for ancestral Y genes or genes from the more ancient neo-Y that derived from the fusion of chromosome XR to the ancestral X (median Ka=0.192 and Ks=0.493; see **Fig. 5**).

Functional evolution of Y genes

Previous work (Koerich *et al.* 2008; Hall *et al.* 2016) and our analysis suggests that many genes on ancestral Y chromosomes were acquired from autosomal locations. The majority of genes on more recently formed neo-Y chromosomes, in contrast, eventually undergo massive degeneration, while some start to diverge early on to be identifiable as male-specific in our pipeline (such as those on the *D. albomicans* or *D. miranda* neo-Y), or are maintained over long periods (such as on the *D. pseudoobscura* neo-Y). To assess which functional pressures are driving the acquisition of new Y genes, or the maintenance or divergence of existing neo-Y genes across flies, we used tissue-specific expression data. On one hand, we assessed expression of putative Y-linked genes with homologs in *D. melanogaster* (see **Fig. 3**) in multiple *D. melanogaster* tissues. We find that most genes that have maintained or acquired Y-linkage are highly expressed in male-specific tissues of *D. melanogaster*, i.e. most genes are highly expressed in testis, and many are also highly expressed in male accessory glands (**Fig. 6A**). To test whether this enrichment for testis- or accessory gland-biased expression is significant, we calculated expression (as TPM; transcripts per million) for all annotated *D. melanogaster* genes (version 6.02) in 5 samples (male head, female head, ovary, testis, accessory glands) and performed binomial tests to evaluate if genes that are Y-linked across Diptera are overrepresented for genes showing highest expression in testis or accessory glands relative to all annotated *D. melanogaster* genes (61 genes out of 106 in our Y-linked gene set vs. 5216 genes out of 17560 genes total; p<0.0001) or whether they are expressed exclusively in testis and accessory glands (36 genes out of 106 in our Y-linked gene set vs. 1655 genes out of 17560 genes total; p<0.0001). A subset of our putative Y-linked genes across Diptera have clear roles in spermatogenesis in *D. melanogaster*. In *Ephydra hians*, for example, a homolog of the *male sterile (2) 34Fe* gene is found on the Y chromosomes, which is highly expressed in male testis, and involved in spermatid differentiation (Lindsley *et al.* 2013); in *T. minor*, a homolog of the *Rcd7* gene is found on the Y, which is involved in spermatogenesis (Lindsley *et al.* 2013); or the *yuri* gene on the Y of *D. miranda* and *D. pseudoobscura*, which is involved in sperm individualization (Texada *et al.* 2008). All these observations are consistent with Y chromosomes being a preferred genomic location for genes with male-specific function (Sturgill *et al.* 2007; Assis *et al.* 2012).

For a subset of species (*D. melanogaster*, *D. albomicans*, *D. miranda*, *D. pseudoobscura*, *E. hians*, *T. dalmanni*, *T. minor*) we had expression data from male and female head, as well as ovary and testis (see **Supplementary Table 5**). This allowed us to compare tissue-specific expression patterns of Y-linked genes directly within a species. Again, we find that most Y genes show highest expression in testis compared to somatic tissue (**Fig. 6B**). This directly demonstrates that surviving or newly acquired Y genes are selected for their male-specific functions. Note that Y-linked genes may show male-specific expression either because their male-specific function makes the (male-limited) Y chromosome an ideal genomic location or because genes on the Y chromosome evolve male-specific functions in response to being located on the Y, and both processes have been found to be important in shaping the gene content of the human Y chromosome (Lahn and Page 1997; Lahn *et al.* 2001). The maintenance of testis-expressed genes on degenerating neo-Y chromosomes (Kaiser *et al.* 2011; Zhou and Bachtrog 2012) and the recruitment of genes to the Y chromosomes whose autosomal paralogs have ancestrally testis-biased expression (as for example inferred from expression patterns in *D. melanogaster*; see **Fig. 6A** or based on testis-biased expression patterns of autosomal or X-linked paralogs of testis-expressed Y-linked transcripts in *D. pseudoobscura*; see **Supplementary Fig. 7**) provides evidence that genes with male-biased expression are selectively acquired or preserved on the Y because of their benefit to males. However, it is possible that some Y-linked genes evolved male-specific expression in response to being located on the male-limited Y chromosome.

Temporal evolution of Y chromosomes

Comparison of tissue-specific expression patterns of Drosophila neo-Y chromosomes reveals an interesting temporal dynamics of Y gene evolution (**Fig. 6B**). On the very recently formed neo-Y chromosome of *D. albomicans*, the majority of genes are not yet differentiated sufficiently to be identified as neo-Y-linked by our pipeline, yet the subset of genes that have accumulated enough mutations so we can pick them up as being located on the Y are predominantly expressed in testis. In *D. miranda*, many more genes on the neo-Y have diverged sufficiently at the DNA sequence level from their neo-X homologs to be identifiable as male-specific, and while many are indeed highly expressed in testis, most are also expressed in somatic (head) tissue. On the older neo-Y of *D. pseudoobscura*, on the other hand, only few genes remain, yet those that have survived are predominantly expressed in testis. This temporal comparison of Y chromosomes paints a dynamic picture of Y gene content evolution, and reveals the importance of male-specific selection shaping Y differentiation (**Fig. 7**). At the earliest stages of Y chromosome formation (as in *D. albomicans*), the majority of genes are indistinguishable on the formerly identical sex chromosomes, and the first genes to diverge at the DNA sequence level on the Y are genes with male-specific function. As time progresses, most genes, independent of their function, start to differentiate and begin to degenerate on the non-recombining Y (as in *D. miranda*). On old Y chromosomes, almost all of the original genes have been lost, and only those with male-specific function will survive on the Y (as in *D. pseudoobscura*), or will be gained secondarily from autosomal paralogs (as in *D. melanogaster*).

Loss of homology between diverging sex chromosomes

The lack of homology between the *D. melanogaster* X and Y chromosome has fueled speculation that the Y in this species is not a degenerate homolog of the X, but instead that the ancestral sex determination system of *Drosophila* was X0, and that the Y was acquired secondarily from a B chromosome (Carvalho 2002). Here, we show that X and Y chromosomes with little homology have evolved independently multiple times in Diptera, and three processes contribute to a lack of homology between X and Y chromosomes (**Fig. 7**). Massive gene loss on the Y is the dominant force shaping sex chromosome divergence, and 100s of genes can quickly erode on a degenerating Y within a few million years. The few genes that are retained on the Y typically have male-specific function, yet exactly those genes are more likely to be lost from the X. In particular, female-biased transmission or the peculiar regulatory mechanisms of the X during spermatogenesis (such as transcriptional suppression of X-linked genes or a lack of dosage compensation in male germline (Vicoso and Bachtrog 2015; Landeen *et al.* 2016); note that the causes of reduced expression of the X chromosome during spermatogenesis are controversial (Vibranovski 2014)) may make it an un-preferred location for testis-expressed genes, and demasculinization (i.e. loss of testis genes) may cause loss of genes on the X that are preferentially maintained on the Y. Finally, recruitment of autosomal genes (typically with male-specific expression) to the Y chromosome means that the closest homologs of many Y genes are located on autosomes.

Thus, our demonstration that Y chromosomes quickly lose homology with the X independently in many lineages with independently formed sex chromosomes and instead acquire genes of autosomal origin argues against the hypothesis that the Y of *D. melanogaster* derives from a supernumerary B chromosome. Furthermore, our comparative analysis in *Drosophila* demonstrates the gradual nature of loss of homology and the various mechanisms contributing to it, and there is thus no need to invoke any additional mechanism (such as a complete loss of the ancestral Y followed by the secondary recruitment of a "B" chromosome) to explain the observed lack of homology between the X and Y of *Drosophila*.

A prominent gene on the Y chromosome in *D. melanogaster*, and in fact the only locus that is shared between the X and Y, is the tandemly repeated rDNA gene family (Ritossa). While there seems to be a general tendency for the rDNA locus to reside on the sex chromosomes in Diptera (Bedo and Webb 1989; Marchi and Pili 1994; Brianti *et al.* 2009), in several species the rDNA is additionally or even exclusively located on autosomes (Stuart *et al.* 1981; Willhoeft 1997; Roy *et al.* 2005). The X and Y rDNA units in *D. melanogaster* are highly similar in sequence due to occasional exchange events (Coen and Dover 1983), and can thus not be detected with our bioinformatics approach that identifies male-specific sequences.

Discussion

The nature of Y chromosomes has remained mysterious. Here, we investigated the gene complement of Y chromosomes in flies, at very different stages of their evolutionary transition. Young neo-Y chromosomes allow us to study gene loss on gene-rich, degenerating Y chromosomes, and the selective forces driving the divergence and maintenance of a subset of genes that were originally present on the Y (Zhou *et al.* 2012; Zhou and Bachtrog 2012; Bachtrog 2013; Zhou and Bachtrog 2015). Comparisons of the ancestral Y chromosome of *Drosophila* species enable us to investigate the dynamics of gene gain and loss on old, homologous Y's (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). Finally, the contrast of old, non-homologous Y chromosomes across Diptera families allows us to identify convergent evolutionary pressures operating on old Y chromosomes.

We find that male-specific selection is a dominating force shaping gene content at each stage of Y evolution. Testis-expressed genes are the first to diverge on very recently formed neo-Y chromosomes (such as in *D. albomicans*), and are preferentially retained during the initial period of massive gene loss on young, degenerating Y chromosomes (such as in *D. miranda* and *D. pseudoobscura*). Once the majority of genes has been lost, Y chromosomes continually reshape their gene complement, by constant losses and gains of genes derived from other locations in the genome with male-specific function (Koerich *et al.* 2008). Additionally, genes ancestrally present on the sex chromosomes with male function may be retained on the Y but lost on the X (as is the case for *D. pseudoobscura*). Thus, after long evolutionary time periods, all homology between the X and Y may be lost. While the 1.5MY old neo-Y of *D. miranda* still shows substantial homology with its former homolog, almost all traces of their shared ancestry have already eroded after 15MY of evolution for the *D. pseudoobscura* Y, and no homology remains between the ancestral sex chromosomes of *Drosophila* (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008).

Independently formed ancient Y chromosomes across flies have evolved similar characteristics convergently: they typically contain very few genes with male-specific function, which appear to be derived mainly from other autosomal locations instead of being remnants of genes ancestrally present on the Y (Carvalho *et al.* 2000; 2001; Carvalho 2002; Koerich *et al.* 2008). The long-term dynamics of ancestral mammalian Y chromosomes is somewhat different. Here, the Y has retained some of its ancestral genes, and they appear to have been maintained for ancestral gene dosage (Cortez *et al.* 2014; Bellott *et al.* 2014). Differences in the mechanism of dosage compensation may contribute to this difference: while male flies generally seem to restore the ancestral gene dosage of X-linked genes through hyper-expression of the X chromosome (Vicoso and Bachtrog 2015), male mammals appear not to globally upregulate X-linked genes (Julien *et al.* 2012; Lin *et al.* 2012). Thus, there may be stronger selection to maintain dosage-sensitive genes on the mammalian X chromosome. Therefore, both lineage-specific as well as general evolutionary mechanisms shape the gene content of Y chromosomes across species.

Methods

Data

We utilized previously published data from separately sequenced male and female genomes for each of the 22 species in our study (Vicoso and Bachtrog 2015). We also sequenced the transcriptomes from male and female whole body separately for each of those species, as described (Vicoso and Bachtrog 2015). We obtained RNA-seq data for male and female heads as well as ovaries and testes for *Drosophila albomicans, D. pseudoobscura, D. miranda, Ephydra hians* and *Themira minor*. Data for the same tissues and male and female whole body for *D. melanogaster* was downloaded from NCBI. Newly collected data have all been uploaded to GenBank. **Supplementary Table 2** gives an overview of all the datasets used, including accession numbers for newly collected sequences.

Coding sequences and protein sequences for *Drosophila melanogaster* genome assembly version r6.2 were downloaded from flybase.org.

Genome assembly

For each species, male and female paired end genomic reads were trimmed and assembled separately using SOAPdenovo (Luo *et al.* 2012) with a kmer size of 31. An overview of the resulting genome assemblies (for females unless otherwise noted) is given in **Supplementary Table 3**.

Transcriptome assembly

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to quality filter the reads. After trimming, Trinity (Grabherr *et al.* 2011) was used to assemble the transcriptomes for each species using default parameters and a kmer size of 25. An overview of the resulting transcriptome assemblies is given in **Supplementary Table 3**.

Pipeline to identify Y-linked coding sequences

We used a subtraction approach to identify putative Y-linked sequences, similar to previous studies done in mammals (Cortez *et al.* 2014; Bellott *et al.* 2014); (see **Supplementary Fig. 1**). Our pipeline starts by making a putative transcriptome assembly of Y-linked genes, and at each step filtering out possible false positives to obtain a conservative list of candidate Y-linked genes. In particular, we first map male RNA-seq reads to the female *de novo* genome assembly using tophat2 (Kim *et al.* 2013) and then build a male *de novo* transcriptome assembly using Trinity (Grabherr *et al.* 2011) from RNA-seq reads that do not map to a female assembly using default parameters and a kmer size of 25. We then mapped the assembled transcripts to the female reference genome using BLAT (Kent 2002) and discarded transcripts if greater than 90% of their length aligned with 98% or greater identity to the female genomic scaffolds or if the blat score was less than 50. Following this, female RNA-seq reads were mapped to the remaining transcripts using bowtie2 (Langmead and Salzberg 2012) with default parameters and transcripts that mapped 50% or more of their sequence with up to two mismatches were discarded. We then did a merging step using the software TGICL (Pertea *et al.* 2003) using a minimum overlap of 30bp and STM, i.e. scaffolding by translational mapping (Surget-Groba and Montoya-Burgos 2010) to remove redundancy and obtain maximal length transcripts. We validated the merged transcripts by mapping them to genomic reads using bowtie2 (Langmead and Salzberg 2012) with default parameters and allowing up to two mismatches. We used soapcoverage (http://soap.genomics.org.cn/soapaligner.html) to calculate male and female genomic coverage for each transcript, and only transcripts for which greater than 60% of their sequence was covered by male reads and less than 10% by female reads were retained. We then mapped male and female RNA-seq reads separately to the remaining transcripts using bowtie2 (Langmead and Salzberg 2012) with default parameters and calculated RPKM values using the software eXpress (Roberts and Pachter 2013). Only transcripts with greater than twice the expression in males compared to females were retained. In order to eliminate transcripts with repetitive sequences, we built repeat libraries for each species using RepARK (Koch *et al.* 2014) and discarded transcripts that mapped to repeats using the software BLAT with default parameters. We then did a final filtering step and discarded transcripts if their effective length used to calculate RPKM as determined by eXpress (Roberts and Pachter 2013) was less than 60% of the total transcript length. We repeated the exact same pipeline but switching sexes in order to identify female-specific transcripts, to empirically assess the false-positive rate of our approach (**Supplementary Table 3**), and only kept species for further analysis where we identified at least twice as many male- relative to female-specific transcripts (see **Supplementary Table 3**). Sequences of all assembled putative Y-linked transcripts are given in **Supplementary Data 1**.

PCR validation for a subset on Y-linked genes

DNA was isolated from two single male and female flies using the Qiagen DNeasy Blood/Tissue kit. PCR primers were designed using the Primer3 software based on assembled putative Y-linked transcripts. PCR amplification was performed with the ThermoFischer Scientific DreamTaq kit, with annealing temperatures ranging from 55˚C to 60˚C.

Finding paralogs and determining Ks values

We calculated Ka, Ks and Ka/Ks values for all Y-linked transcripts for which we could find paralogous sequences in the female genome assembly. To this end, we first determined putative peptide sequences for the candidate Y-linked transcripts using either CPC (Kong *et al.* 2007) or ORF finder (http://www.bioinformatics.org/sms2/orf_find.html). We then used tblastn (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to map putative Y-linked peptides to the female genome assembly for each species, to identify whether Y-linked transcripts had paralogous sequences in the female genome. We ignored all transcripts that aligned poorly with a BLAST score below 50 or with less than 40% sequence identity. We then used the software exonerate (Slater and Birney 2005) with parameters : --exhaustive, protein2genome, -n 1, to extract coding sequences for the Y-linked transcripts as well as their paralogs in the female genomes and then aligned Y-linked transcripts to the coding sequences of these paralogs using the software prank (http://wasabiapp.org/software/prank). Finally, we used KaKs_Calculator (Zhang *et al.* 2006) to determine Ka, Ks and Ka/Ks values.

For the four species in our analysis with neo-sex chromosomes, we used homology to *D. melanogaster* as well as their published genome assemblies to determine the chromosomal location of paralogs of putative Y-linked transcripts. To this end, we mapped female genomic scaffolds to coding sequences from *D. melanogaster* and the published species genomes using BLAT with default parameters and then used the best alignment to assign paralogs to chromosomal arms. We classified transcripts as being autosomal, neo-X-linked, X-linked or ancestrally Y-linked (i.e. genes that are Y-linked in *D. melanogaster* or in *D. pseudoobscura,* for *D. miranda*) based on the chromosomal location that we determined for their paralogs. Paralogs whose chromosomal location could not be identified were placed in the 'unknown' category.

Coverage analysis for paralogs of putative Y-linked transcripts

We used previously published genomic coverage data as well as genome assemblies (Vicoso and Bachtrog 2015) to determine the coverage of the genomic scaffolds that the paralogs of the putative Y-linked transcripts in the female genome are located on. We then plotted a histogram of log2 (Normalized Male/Female) coverage for all genomic scaffolds highlighting the coverage of the scaffolds containing the Y-linked paralog in red lines (see **Figure 4**).

For *D. busckii,* no published coverage data was available. We used SOAPdenovo to build a genome assembly from male and female genomic reads and then aligned male and female genomic reads separately to the *de novo* assembled genome using bowtie2 (Langmead and Salzberg 2012)with default parameters. We then used soapcoverage to calculate male and female genomic coverages for all scaffolds whose length was at least 1000bp. We then proceeded similarly to the other species in the analysis to obtain a coverage histogram.

Tissue-specific expression

For the six species in our analysis for which we had RNA-seq data from male and female heads, ovaries and testes (see **Supplementary Table 2**), we calculated expression of the Y-linked transcripts for each tissue as TPM (transcripts per million) values using the software kallisto (Bray *et al.* 2016) with default parameters.

Tissue-specific expression of Y homologs in *D. melanogaster*

For each species, we used BLAT (Kent 2002) with a translated nucleotide and a translated database to identify the *D. melanogaster* genes that are homologous to the putative Y-linked transcripts using default parameters and a BLAT score cutoff of 50. We then downloaded tissue-specific expression profiles for each of those genes from flybase.org, in order to investigate the spatio-temporal expression patterns of genes in *D. melanogaster* whose homologs have become Y-linked in the different fly species investigated.

Data availability

Newly collected data have all been uploaded to GenBank. **Supplementary Table 2** gives an overview of all the datasets used, including accession numbers for newly collected sequences (Bioproject PRJNA385725 [SRX2822436- SRX2822453] and SRX2788162).

.

References

Assis R., Zhou Q., Bachtrog D., 2012 Sex-biased transcriptome evolution in Drosophila. Genome Biol Evol **4**: 1189–1200.

Bachtrog D., 2013 Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. **14**: 113–124.

Bachtrog D., Mank J. E., Peichel C. L., Kirkpatrick M., Otto S. P., Ashman T.-L., Hahn M. W., Kitano J., Mayrose I., Ming R., Perrin N., Ross L., Valenzuela N., Vamosi J. C., Tree of Sex Consortium, 2014 Sex determination: why so many ways of doing it? PLoS Biol. **12**: e1001899.

Bedo D. G., Webb G. C., 1989 Conservation of nucleolar structure in polytene tissues of Ceratitis capitata (Diptera: Tephritidae). Chromosoma **98**: 443–449.

Bellott D. W., Hughes J. F., Skaletsky H., Brown L. G., Pyntikova T., Cho T.-J., Koutseva N., Zaghlul S., Graves T., Rock S., Kremitzki C., Fulton R. S., Dugan S., Ding Y., Morton D., Khan Z., Lewis L., Buhay C., Wang Q., Watt J., Holder M., Lee S., Nazareth L., Rozen S., Muzny D. M., Warren W. C., Gibbs R. A., Wilson R. K., Page D. C., 2014 Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature **508**: 494–499.

Blackmon H., Ross L., Bachtrog D., 2017 Sex Determination, Sex Chromosomes, and Karyotype Evolution in Insects. J. Hered. **108**: 78–93.

Bray N. L., Pimentel H., Melsted P., Pachter L., 2016 Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**: 525–527.

Brianti M. T., Ananina G., Recco-Pimentel S. M., Klaczko L. B., 2009 Comparative analysis of the chromosomal positions of rDNA genes in species of the tripunctata radiation of Drosophila. Cytogenet. Genome Res. **125**: 149–157.

Carvalho A. B., 2002 Origin and evolution of the Drosophila Y chromosome. Curr. Opin. Genet. Dev. **12**: 664–668.

Carvalho A. B., Clark A. G., 2005 Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. Science **307**: 108–110.

Carvalho A. B., Clark A. G., 2013 Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome Res. **23**: 1894–1907.

Carvalho A. B., Dobo B. A., Vibranovski M. D., Clark A. G., 2001 Identification of five new genes on the Y chromosome of Drosophila melanogaster. Proc. Natl. Acad. Sci. U.S.A.

98: 13225–13230.

Carvalho A. B., Lazzaro B. P., Clark A. G., 2000 Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. Proc. Natl. Acad. Sci. U.S.A. **97**: 13239–13244.

Carvalho A. B., Vicoso B., Russo C. A. M., Swenor B., Clark A. G., 2015 Birth of a new gene on the Y chromosome of Drosophila melanogaster. Proc. Natl. Acad. Sci. U.S.A. **112**: 12450–12455.

Charlesworth B., Charlesworth D., 2000 The degeneration of Y chromosomes. Philos. Trans. R. Soc. Lond., B, Biol. Sci. **355**: 1563–1572.

Coen E. S., Dover G. A., 1983 Unequal exchanges and the coevolution of X and Y rDNA arrays in Drosophila melanogaster. Cell **33**: 849–855.

Cortez D., Marin R., Toledo-Flores D., Froidevaux L., Liechti A., Waters P. D., Grützner F., Kaessmann H., 2014 Origins and functional evolution of Y chromosomes across mammals. Nature **508**: 488–493.

Daines B., Wang H., Wang L., Li Y., Han Y., Emmert D., Gelbart W., Wang X., Li W., Gibbs R., Chen R., 2011 The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res. **21**: 315–324.

Gatti M., Pimpinelli S., 1992 Functional elements in Drosophila melanogaster heterochromatin. Annu. Rev. Genet. **26**: 239–275.

Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B. W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A., 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**: 644–652.

Hall A. B., Papathanos P.-A., Sharma A., Cheng C., Akbari O. S., Assour L., Bergman N. H., Cagnetti A., Crisanti A., Dottorini T., Fiorentini E., Galizi R., Hnath J., Jiang X., Koren S., Nolan T., Radune D., Sharakhova M. V., Steele A., Timoshevskiy V. A., Windbichler N., Zhang S., Hahn M. W., Phillippy A. M., Emrich S. J., Sharakhov I. V., Tu Z. J., Besansky N. J., 2016 Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. Proc. Natl. Acad. Sci. U.S.A. **113**: E2114–23.

Hall A. B., Qi Y., Timoshevskiy V., Sharakhova M. V., Sharakhov I. V., Tu Z., 2013 Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. BMC Genomics **14**: 273.

Holt R. A., Subramanian G. M., Halpern A., Sutton G. G., Charlab R., Nusskern D. R., Wincker P., Clark A. G., Ribeiro J. M. C., Wides R., Salzberg S. L., Loftus B., Yandell M., Majoros W. H., Rusch D. B., Lai Z., Kraft C. L., Abril J. F., Anthouard V., Arensburger P., Atkinson P. W., Baden H., de Berardinis V., Baldwin D., Benes V., Biedler J., Blass C., Bolanos R., Boscus D., Barnstead M., Cai S., Center A., Chaturverdi K., Christophides G. K., Chrystal M. A., Clamp M., Cravchik A., Curwen V., Dana A., Delcher A., Dew I., Evans C. A., Flanigan M., Grundschober-Freimoser A., Friedli L., Gu Z., Guan P., Guigo R., Hillenmeyer M. E., Hladun S. L., Hogan J. R., Hong Y. S., Hoover J., Jaillon O., Ke Z., Kodira C., Kokoza E., Koutsos A., Letunic I., Levitsky A., Liang Y., Lin J.-J., Lobo N. F., Lopez J. R., Malek J. A., McIntosh T. C., Meister S., Miller J., Mobarry C., Mongin E., Murphy S. D., O'Brochta D. A., Pfannkoch C., Qi R., Regier M. A., Remington K., Shao H., Sharakhova M. V., Sitter C. D., Shetty J., Smith T. J., Strong R., Sun J., Thomasova D., Ton L. Q., Topalis P., Tu Z., Unger M. F., Walenz B., Wang A., Wang J., Wang M., Wang X., Woodford K. J., Wortman J. R., Wu M., Yao A., Zdobnov E. M., Zhang H., Zhao Q., Zhao S., Zhu S. C., Zhimulev I., Coluzzi M., Torre della A., Roth C. W., Louis C., Kalush F., Mural R. J., Myers E. W., Adams M. D., Smith H. O., Broder S., Gardner M. J., Fraser C. M., Birney E., Bork P., Brey P. T., Venter J. C., Weissenbach J., Kafatos F. C., Collins F. H., Hoffman S. L., 2002 The genome sequence of the malaria mosquito Anopheles gambiae. Science **298**: 129–149.

Hughes J. F., Page D. C., 2015 The Biology and Evolution of Mammalian Y Chromosomes. Annu. Rev. Genet. **49**: 507–527.

Hughes J. F., Skaletsky H., Pyntikova T., Graves T. A., van Daalen S. K. M., Minx P. J., Fulton R. S., McGrath S. D., Locke D. P., Friedman C., Trask B. J., Mardis E. R., Warren W. C., Repping S., Rozen S., Wilson R. K., Page D. C., 2010 Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature **463**: 536–539.

Julien P., Brawand D., Soumillon M., Necsulea A., Liechti A., Schütz F., Daish T., Grützner F., Kaessmann H., 2012 Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. (NH Barton, Ed.). PLoS Biol. **10**: e1001328.

Kaiser V. B., Zhou Q., Bachtrog D., 2011 Nonrandom gene loss from the Drosophila miranda neo-Y chromosome. Genome Biol Evol **3**: 1329–1337.

Kent W. J., 2002 BLAT--the BLAST-like alignment tool. Genome Res. **12**: 656–664.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S. L., 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**: R36.

Koch P., Platzer M., Downie B. R., 2014 RepARK--de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. **42**: e80–e80.

Koerich L. B., Wang X., Clark A. G., Carvalho A. B., 2008 Low conservation of gene content in the Drosophila Y chromosome. Nature **456**: 949–951.

Kong L., Zhang Y., Ye Z.-Q., Liu X.-Q., Zhao S.-Q., Wei L., Gao G., 2007 CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. **35**: W345–9.

Krsticevic F. J., Santos H. L., Januário S., Schrago C. G., Carvalho A. B., 2010 Functional copies of the Mst77F gene on the Y chromosome of Drosophila melanogaster. Genetics **184**: 295–307.

Krsticevic F. J., Schrago C. G., Carvalho A. B., 2015 Long-Read Single Molecule Sequencing to Resolve Tandem Gene Copies: The Mst77Y Region on the Drosophila melanogaster Y Chromosome. G3 (Bethesda) **5**: 1145–1150.

Krzywinska E., Dennison N. J., Lycett G. J., Krzywinski J., 2016 A maleness gene in the malaria mosquito Anopheles gambiae. Science **353**: 67–69.

Lahn B. T., Page D. C., 1997 Functional coherence of the human Y chromosome. Science **278**: 675–680.

Lahn B. T., Pearson N. M., Jegalian K., 2001 The human Y chromosome, in the light of evolution. Nat. Rev. Genet. **2**: 207–216.

Landeen E. L., Muirhead C. A., Wright L., Meiklejohn C. D., Presgraves D. C., 2016 Sex Chromosome-wide Transcriptional Suppression and Compensatory Cis-Regulatory Evolution Mediate Gene Expression in the Drosophila Male Germline. (PB Becker, Ed.). PLoS Biol. **14**: e1002499.

Langmead B., Salzberg S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

Larracuente A. M., Noor M. A. F., Clark A. G., 2010 Translocation of Y-linked genes to the dot chromosome in Drosophila pseudoobscura. Mol. Biol. Evol. **27**: 1612–1620.

Lin F., Xing K., Zhang J., He X., 2012 Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. Proc. Natl. Acad. Sci. U.S.A. **109**: 11752–11757.

Lindsley D. L., Roote J., Kennison J. A., 2013 Anent the genomics of spermatogenesis in Drosophila melanogaster. (P Callaerts, Ed.). PLoS ONE **8**: e55915.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D. W., Yiu S.-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.-W., Wang J., 2012

SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1**: 18.

Marchi A., Pili E., 1994 Ribosomal RNA genes in mosquitoes: localization by fluorescence in situ hybridization (FISH). Heredity (Edinb) **72 (Pt 6)**: 599–605.

Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J., Quackenbush J., 2003 TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19**: 651–652.

Ritossa F., The *bobbed* locus. In:Ashburner M, Novitski E (Eds.), *Genetics and Biology of Drosophila*,

Roberts A., Pachter L., 2013 Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods **10**: 71–73.

Roy V., Monti-Dedieu L., Chaminade N., Siljak-Yakovlev S., Aulard S., Lemeunier F., Montchamp-Moreau C., 2005 Evolution of the chromosomal location of rDNA genes in two Drosophila species subgroups: ananassae and melanogaster. Heredity (Edinb) **94**: 388–395.

Skaletsky H., Kuroda-Kawaguchi T., Minx P. J., Cordum H. S., Hillier L., Brown L. G., Repping S., Pyntikova T., Ali J., Bieri T., Chinwalla A., Delehaunty A., Delehaunty K., Du H., Fewell G., Fulton L., Fulton R., Graves T., Hou S.-F., Latrielle P., Leonard S., Mardis E., Maupin R., McPherson J., Miner T., Nash W., Nguyen C., Ozersky P., Pepin K., Rock S., Rohlfing T., Scott K., Schultz B., Strong C., Tin-Wollam A., Yang S.-P., Waterston R. H., Wilson R. K., Rozen S., Page D. C., 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**: 825–837.

Slater G. S. C., Birney E., 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**: 31.

Soh Y. Q. S., Alföldi J., Pyntikova T., Brown L. G., Graves T., Minx P. J., Fulton R. S., Kremitzki C., Koutseva N., Mueller J. L., Rozen S., Hughes J. F., Owens E., Womack J. E., Murphy W. J., Cao Q., de Jong P., Warren W. C., Wilson R. K., Skaletsky H., Page D. C., 2014 Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell **159**: 800–813.

Stuart W. D., Bishop J. G., Carson H. L., Frank M. B., 1981 Location of the 18/28S ribosomal RNA genes in two Hawaiian Drosophila species by monoclonal immunological identification of RNA.DNA hybrids in situ. Proc. Natl. Acad. Sci. U.S.A. **78**: 3751–3754.

Sturgill D., Zhang Y., Parisi M., Oliver B., 2007 Demasculinization of X chromosomes in the

Drosophila genus. Nature **450**: 238–241.

Surget-Groba Y., Montoya-Burgos J. I., 2010 Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. **20**: 1432–1440.

Texada M. J., Simonette R. A., Johnson C. B., Deery W. J., Beckingham K. M., 2008 Yuri gagarin is required for actin, tubulin and basal body functions in Drosophila spermatogenesis. J. Cell. Sci. **121**: 1926–1936.

Vibranovski M. D., 2014 Meiotic sex chromosome inactivation in Drosophila. Journal of Genomics **2**: 104–117.

Vicoso B., Bachtrog D., 2015 Numerous transitions of sex chromosomes in Diptera. (HS Malik, Ed.). PLoS Biol. **13**: e1002078.

Vicoso B., Charlesworth B., 2006 Evolution on the X chromosome: unusual patterns and processes. Nat. Rev. Genet. **7**: 645–653.

White M. J. D., 1949 Cytological Evidence on the Phylogeny and Classification of the Diptera. Evolution **3**: 252.

Willhoeft U., 1997 Fluorescence in situ hybridization of ribosomal DNA to mitotic chromosomes of tsetse flies (Diptera: Glossinidae: Glossina). Chromosome Res. **5**: 262–267.

Zhang Z., Li J., Zhao X.-Q., Wang J., Wong G. K.-S., Yu J., 2006 KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics **4**: 259–263.

Zhou Q., Bachtrog D., 2012 Sex-specific adaptation drives early sex chromosome evolution in Drosophila. Science **337**: 341–345.

Zhou Q., Bachtrog D., 2015 Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in Drosophila. (A Akhtar, Ed.). PLoS Genet. **11**: e1005331.

Zhou Q., Zhu H.-M., Huang Q.-F., Zhao L., Zhang G.-J., Roy S. W., Vicoso B., Xuan Z.-L., Ruan J., Zhang Y., Zhao R.-P., Ye C., Zhang X.-Q., Wang J., Wang W., Bachtrog D., 2012 Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. BMC Genomics **13**: 109.

Acknowledgements Funded by NIH grants (R01GM076007, GM101255 and R01GM093182) to DB. We thank Beatriz Vicoso for helpful discussions on this project and Zaak Walton, Matthew Nalley and Alison Nguyen for technical assistance.

Author contributions SM and DB conceived the study, collected and analyzed the data and wrote the manuscript.

Figures

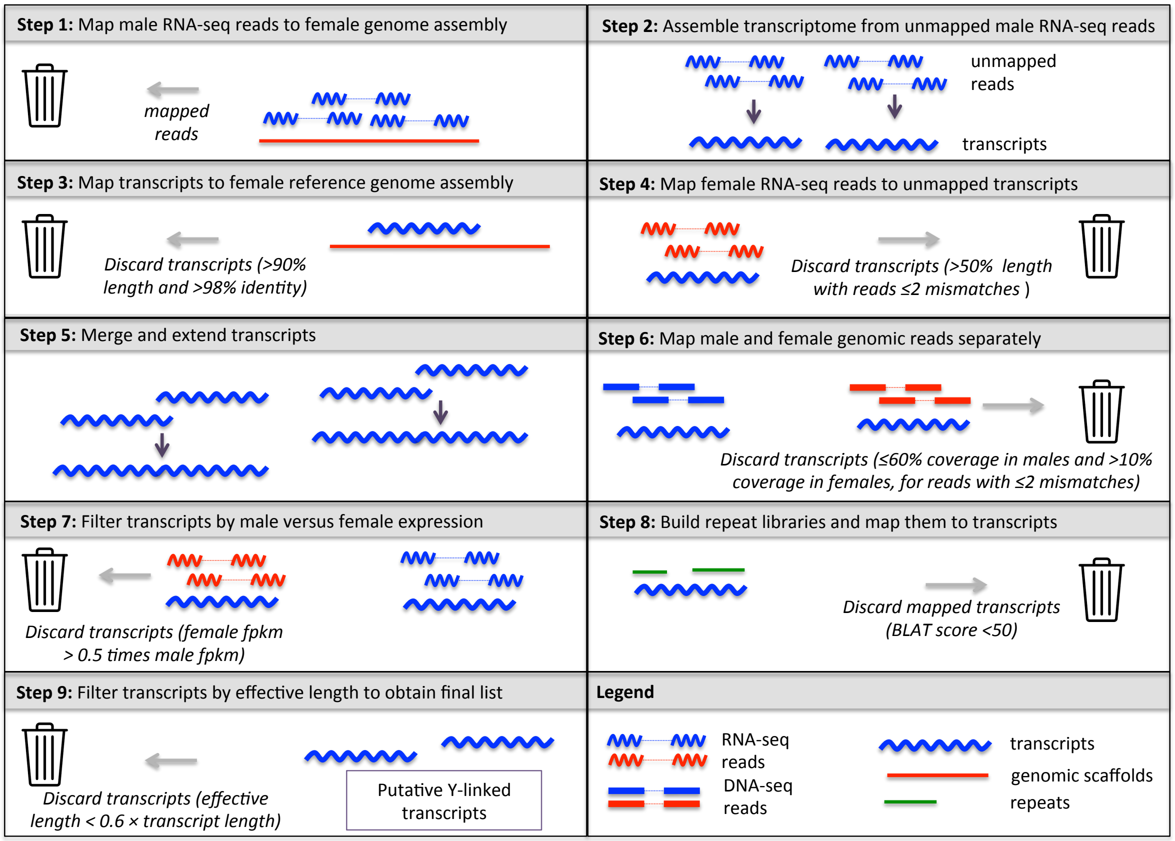

Step 1: Map male RNA-seq reads to female genome assembly
mapped reads
Step 2: Assemble transcriptome from unmapped male RNA-seq reads
unmapped reads
transcripts
Step 3: Map transcripts to female reference genome assembly
Discard transcripts (>90% length and >98% identity)
Step 4: Map female RNA-seq reads to unmapped transcripts
Discard transcripts (>50% length with reads ≤2 mismatches)
Step 5: Merge and extend transcripts
Step 6: Map male and female genomic reads separately
Discard transcripts (≤60% coverage in males and >10% coverage in females, for reads with ≤2 mismatches)
Step 7: Filter transcripts by male versus female expression
Discard transcripts (female fpkm > 0.5 times male fpkm)
Step 8: Build repeat libraries and map them to transcripts
Discard mapped transcripts (BLAT score <50)
Step 9: Filter transcripts by effective length to obtain final list
Discard transcripts (effective length < 0.6 × transcript length)
Putative Y-linked transcripts
Legend
RNA-seq reads
DNA-seq reads
transcripts
genomic scaffolds
repeats

Figure 1. Bioinformatic subtraction pipeline to infer Y-linked transcripts.

Male RNA-seq reads are mapped to genomic scaffolds build from female genomic reads (Step 1); unmapped male RNA-seq reads are used to build a *de novo* transcriptome (Step 2), and transcripts that either map to the female genome assembly (Step 3) or female RNA-seq reads (Step 4) are discarded. Remaining transcripts are merged (Step 5) and only merged transcripts are kept that show mapping to male genomic reads and no mapping to female genomic reads (Step 6) and that show expression in males but not females (Step 7). Transcripts that mapped to a *de novo* repeat library were discarded (Step 8), and only transcripts which had an effective length (as calculated by the software eXpress) greater than 0.6 times the transcript length were kept in the final list (Step 9).

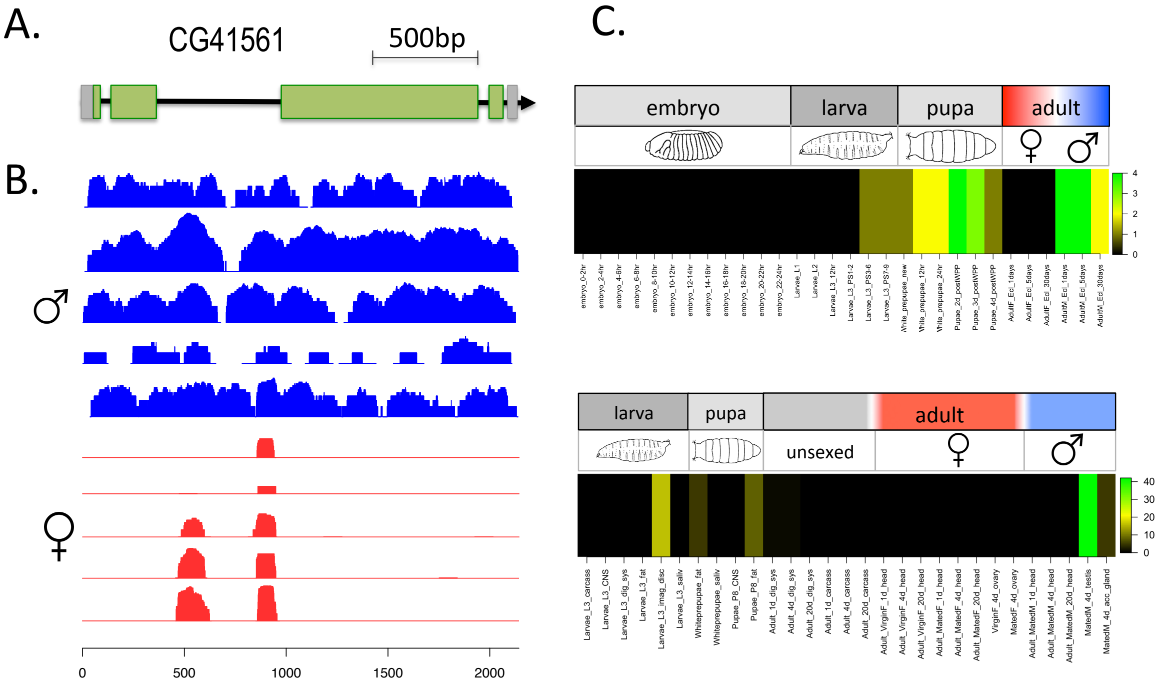
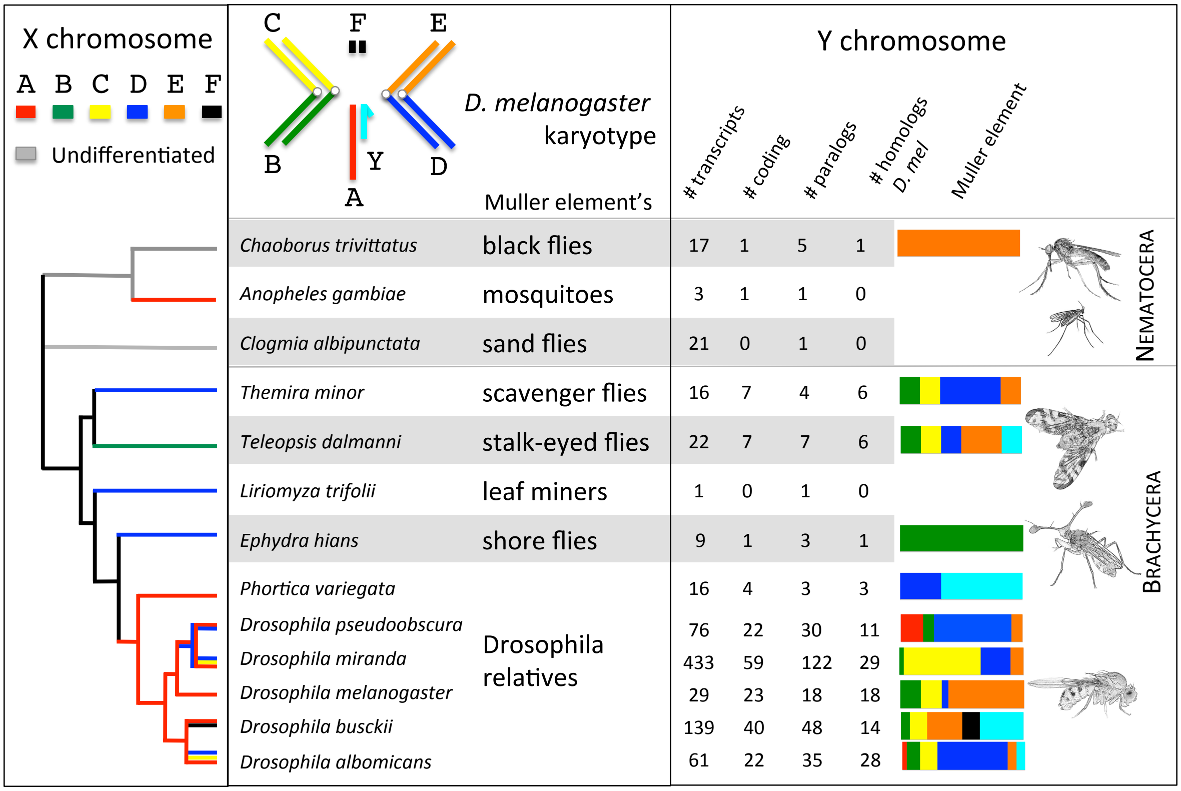


Figure 2. ***CG41561*****, a new protein-coding gene on the *D. melanogaster* Y chromosome**.

A. Intron/exon structure of *CG41561* (grey are non-coding exons, green are coding exons). **B**. Mapping of five male (blue) and five female (red) Drosophila genomic reads to *CG41561* (for strain information see **Supplementary Table 2**). **C**. Expression profile of *CG41561* across developmental stages (top; samples are ordered by developmental time) and larval and adult tissues (bottom); colors in heatmap refer to expression level. *CG41561* is first expressed in third instar larvae, and shows maximum expression in pupae and young adult males (it is not expressed in females). Across tissues, *CG41561* is expressed in imaginal discs of third instar larvae and most highly in adult male testis. Expression profiles are taken from flybase. Abbreviations: CNS: central nervous system, dig_sys: digestive system, fat: fat body, imag_disc: imaginal disc, saliv: salivary gland, acc_gland: accessory gland, 1d: 1-day, 4d: 4-days, 20d: 20-days.



Species	Family	# transcripts	# coding	# paralogs	# homologs D. mel
Chaoborus trivittatus	black flies	17	1	5	1
Anopheles gambiae	mosquitoes	3	1	1	0
Clogmia albipunctata	sand flies	21	0	1	0
Themira minor	scavenger flies	16	7	4	6
Teleopsis dalmanni	stalk-eyed flies	22	7	7	6
Liriomyza trifolii	leaf miners	1	0	1	0
Ephydra hians	shore flies	9	1	3	1
Phortica variegata	Drosophila relatives	16	4	3	3
Drosophila pseudoobscura		76	22	30	11
Drosophila miranda		433	59	122	29
Drosophila melanogaster		29	23	18	18
Drosophila busckii		139	40	48	14
Drosophila albomicans		61	22	35	28

Figure 3. Y gene content evolution in flies.

Shown are the species for which we have identified Y-linked transcripts. The karyotype of *D. melanogaster* males is shown (consisting of Muller element's A-F), and the color-coded branches of the phylogeny indicate which chromosome arm (Muller element) segregates as the sex chromosome in a species (from ref.). The table gives the number of Y-linked transcripts identified from each species (#transcripts), the number of Y-linked transcripts that are predicted to be protein-coding (#coding), the number of Y-linked transcripts for which we could detect a paralog within the female genome of a particular species (#paralogs), and the number of Y-linked transcripts for which we could detect a homolog in *D. melanogaster* (#homologs *D. mel*). The bar charts indicate to which Muller element Y-linked homologs match within the *D. melanogaster* genome (with Y-linked genes of *D. melanogaster* being shown in turquoise); for *D. melanogaster*, mapping of Y-linked genes within the female *D. melanogaster* genome is shown. Shading distinguishes between different Diptera families (indicated by the common name of that family).

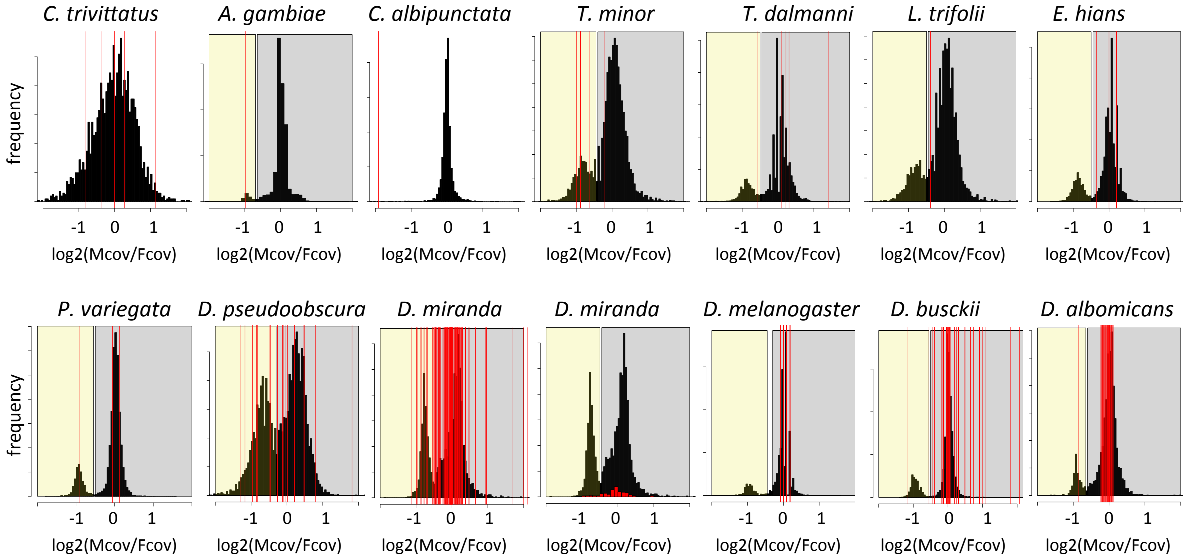


Figure 4. Genomic coverage of paralogs of putative Y-linked genes.

The histograms show the male/female genomic coverage ratio of all scaffolds in the female genome (> 1000bp). Scaffolds that are X-linked have reduced male/female coverage ratio (log2 (Mcov/Fcov) around -1), while autosomal scaffolds have similar coverage in males and females (log2 (Mcov/Fcov) around 0). The red lines indicate the male/female coverage ratio of scaffolds that have homologs on the Y chromosomes. Y-linked genes with X-linked homologs should show half the male/female coverage compared to ones with autosomal paralogs. Putative X-linked scaffolds are indicated by their yellow shading, and putative autosomal scaffolds are shown in grey shading. Note that we also show a histogram for paralogs of *D. miranda* Y-linked transcripts, due to the large number of Y-linked transcripts.

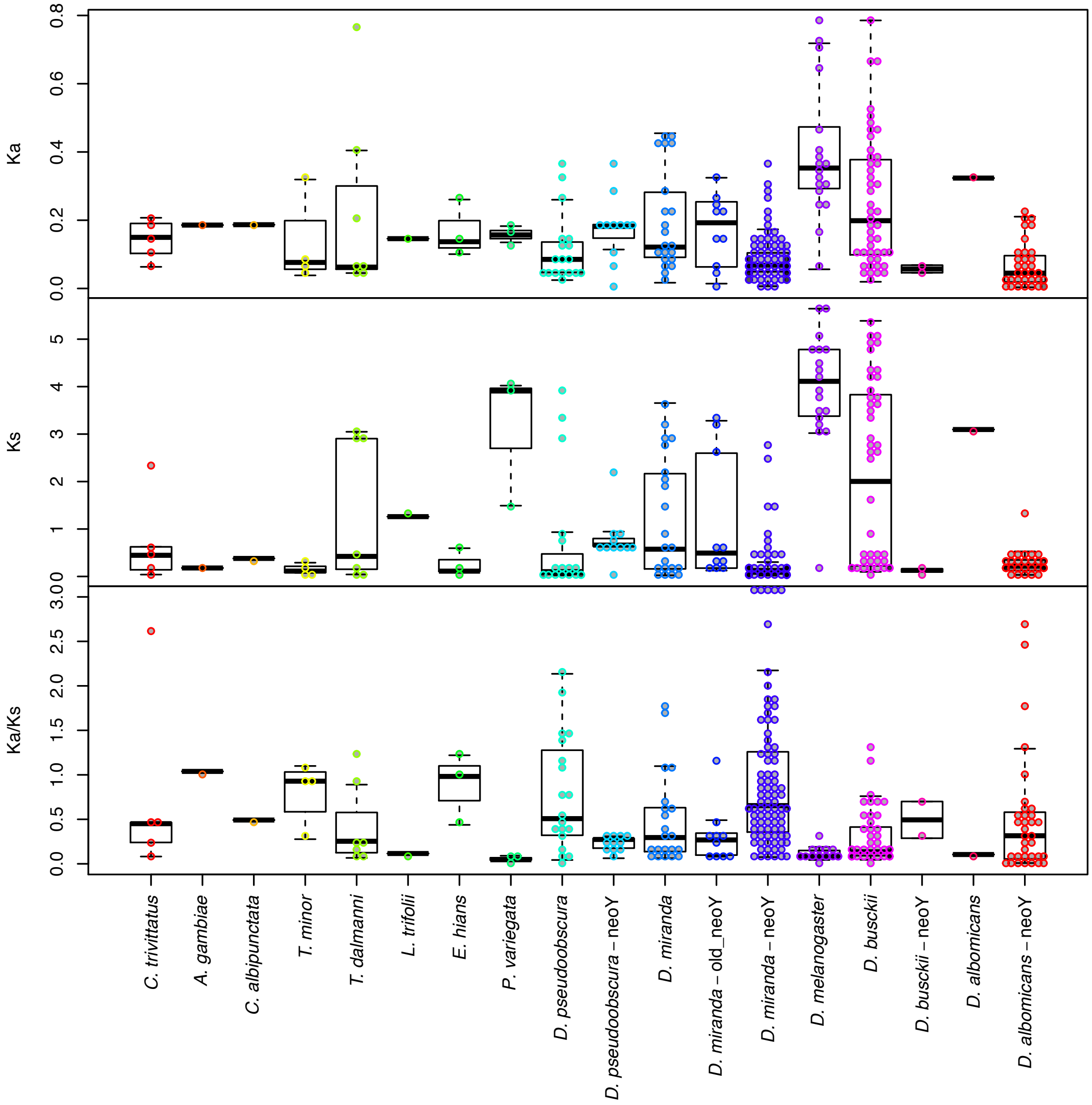


Figure 5. Divergence analysis of Y linked genes.

Shown are rates of sequence evolution of Y-linked genes and their closest paralogs within the female genome. For *Drosophila* species with neo-sex chromosomes, we show divergence between neo-X/neo-Y homologs (i.e. Y-linked genes with their closest homolog on the neo-X) separately from other Y-linked genes. Shown are rates of amino acid evolution (Ka), rates of synonymous evolution (Ks), and their ratio (Ka/Ks). Note that Ka/Ks values > 3 are plotted at 3.

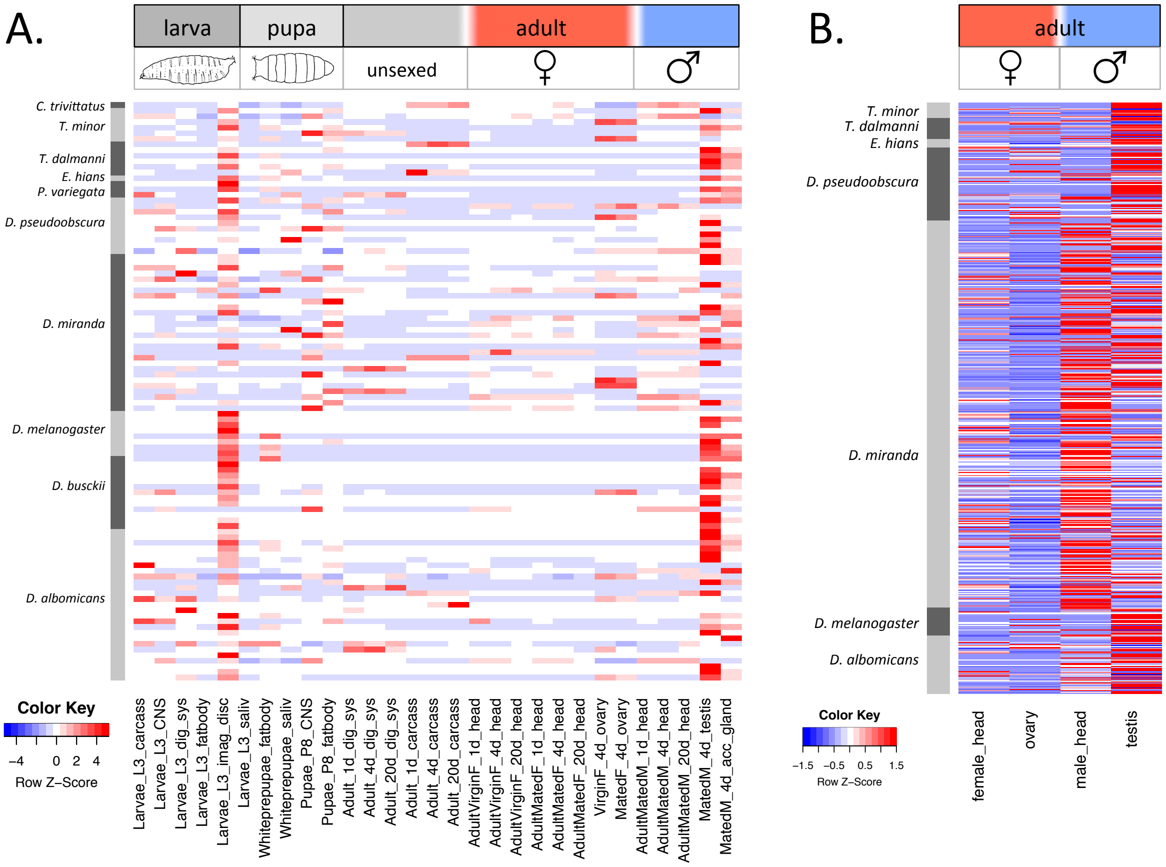


Figure 6. Functional specialization of Y-linked genes in flies.

A. Expression patterns of putative Y-linked genes with homologs in *D. melanogaster* in multiple *D. melanogaster* tissues. **B.** Expression patterns of putative Y-linked genes in male and female head, and testis and ovaries for different species. Expression values were calculated as TPM (transcript per million) and row normalized to obtain z-scores with a mean of 0 and standard deviation of 1, using the built-in scale='row' argument in the heatmap.2 function from the package gplots in R.

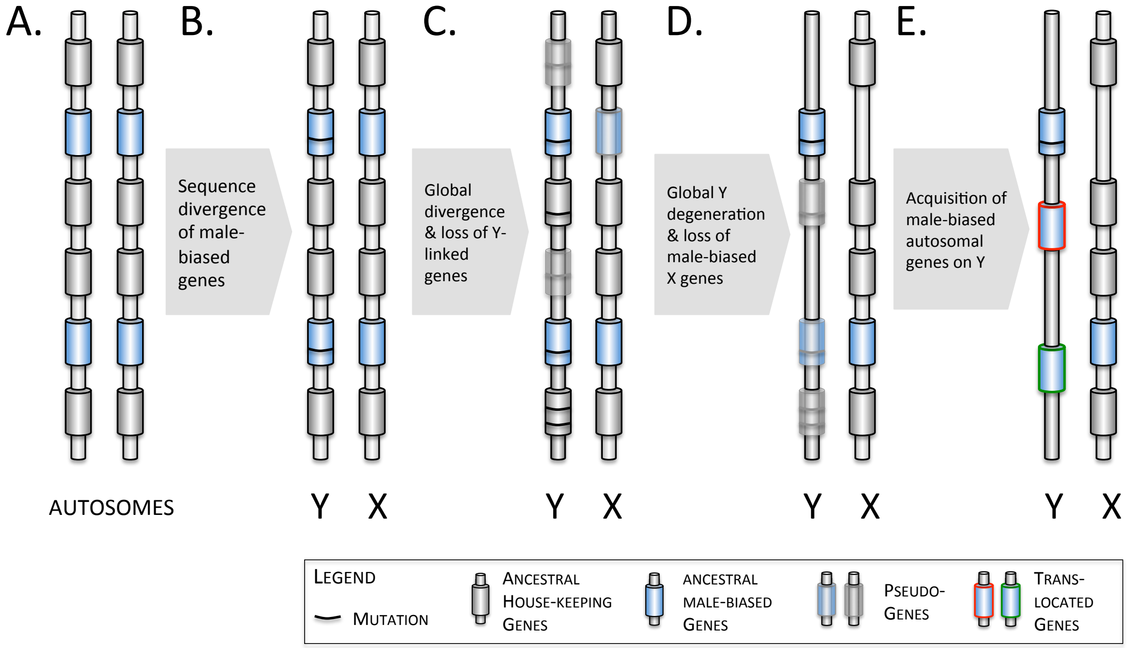
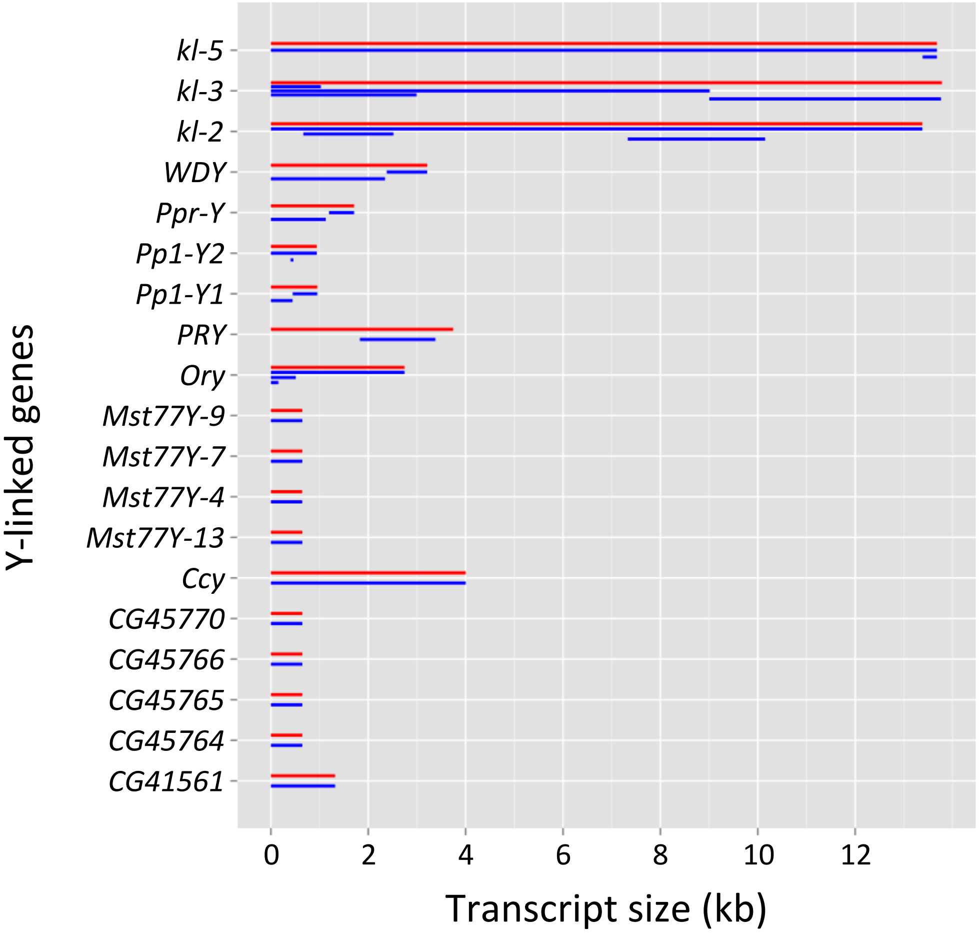

A.
B.
C.
D.
E.
Sequence divergence of male-biased genes
Global divergence & loss of Y-linked genes
Global Y degeneration & loss of male-biased X genes
Acquisition of male-biased autosomal genes on Y
AUTOSOMES
Y
X
Y
X
Y
X
Y
X
LEGEND
MUTATION
ANCESTRAL HOUSE-KEEPING GENES
ANCESTRAL MALE-BIASED GENES
PSEUDO-GENES
TRANS-LOCATED GENES

Figure 7. Model of Y-linked gene content evolution in flies.

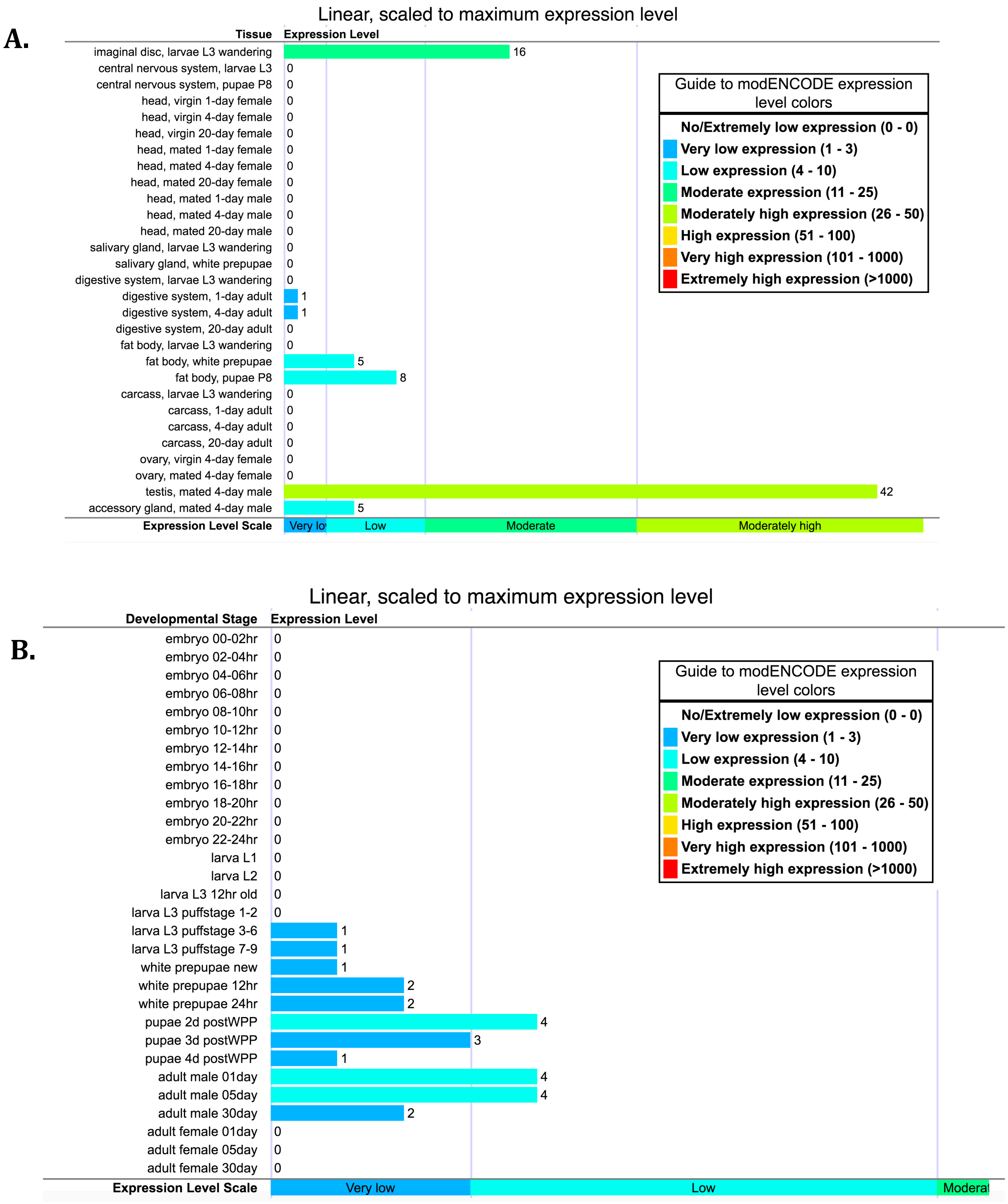
The dynamics of gene content evolution in flies across time is shown. (A) X and Y chromosomes originate from ordinary autosomes with identical gene content. (B) The first genes to diverge at the DNA sequence level are genes with male-biased expression. (C) Over time, most genes on the Y acquire mutations, and many start to become pseudogenes. (D) Continuing Y degeneration, and loss of some male-biased genes on the X chromosome. (E). Acquisition of male-biased genes from autosomes to the Y chromosome. Genes with male-biased expression are shown by blue shading, and genes with broad functions are shown with grey shading.

Supplementary Figures



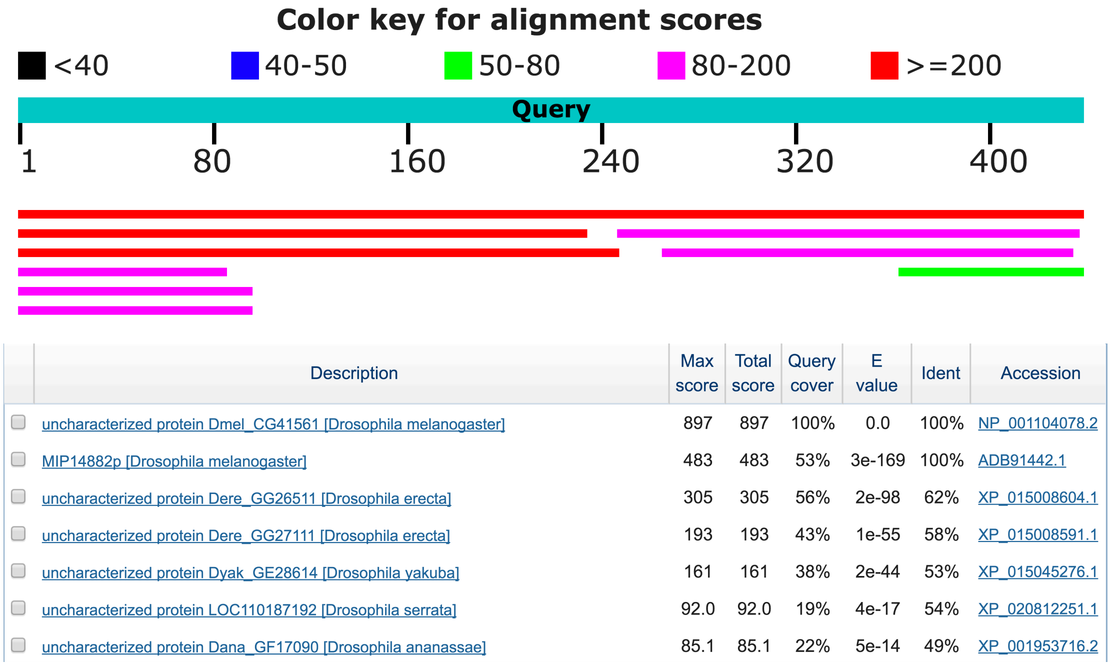
Supplementary Figure 1. Validation of bioinformatics pipeline to infer Y-linked transcripts in *D. melanogaster.*

All annotated transcripts of the *D. melanogaster* Y chromosome are shown as red lines. Blue lines indicate the alignment of assembled transcripts using our subtraction pipeline (from Fig. 1) against known *D. melanogaster* Y genes. The majority of genes on the *D. melanogaster* Y chromosome are covered by a single transcript assembled by our pipeline. For some genes, multiple isoforms were assembled by our pipeline (overlapping blue lines).



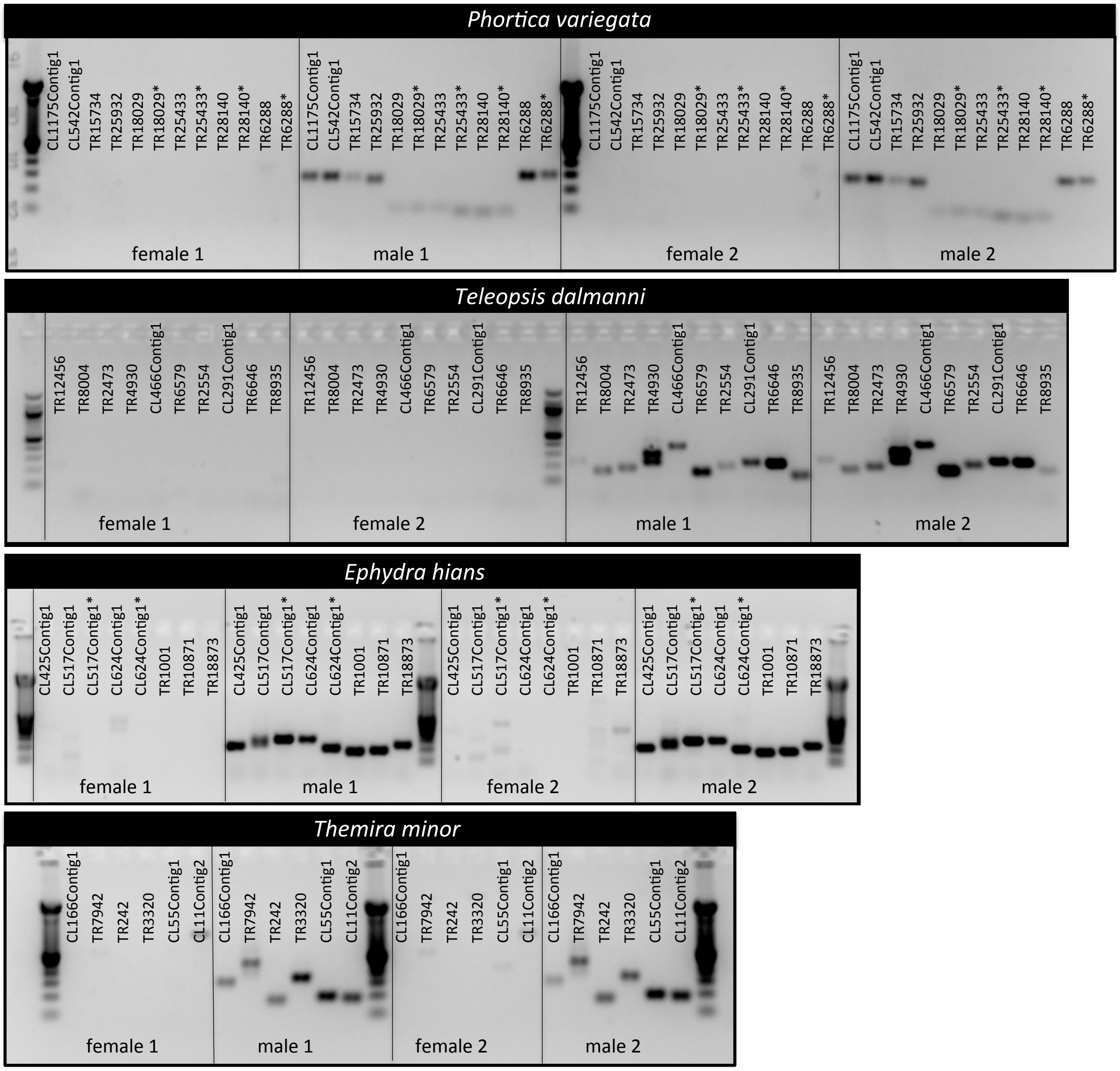
Supplementary Figure 2. Expression profile of the new Y-linked gene *CG41561*.

(**a**) Tissue expression and (**b**) developmental stage expression for *CG41561*/transcript TR3794 (images taken from flybase).



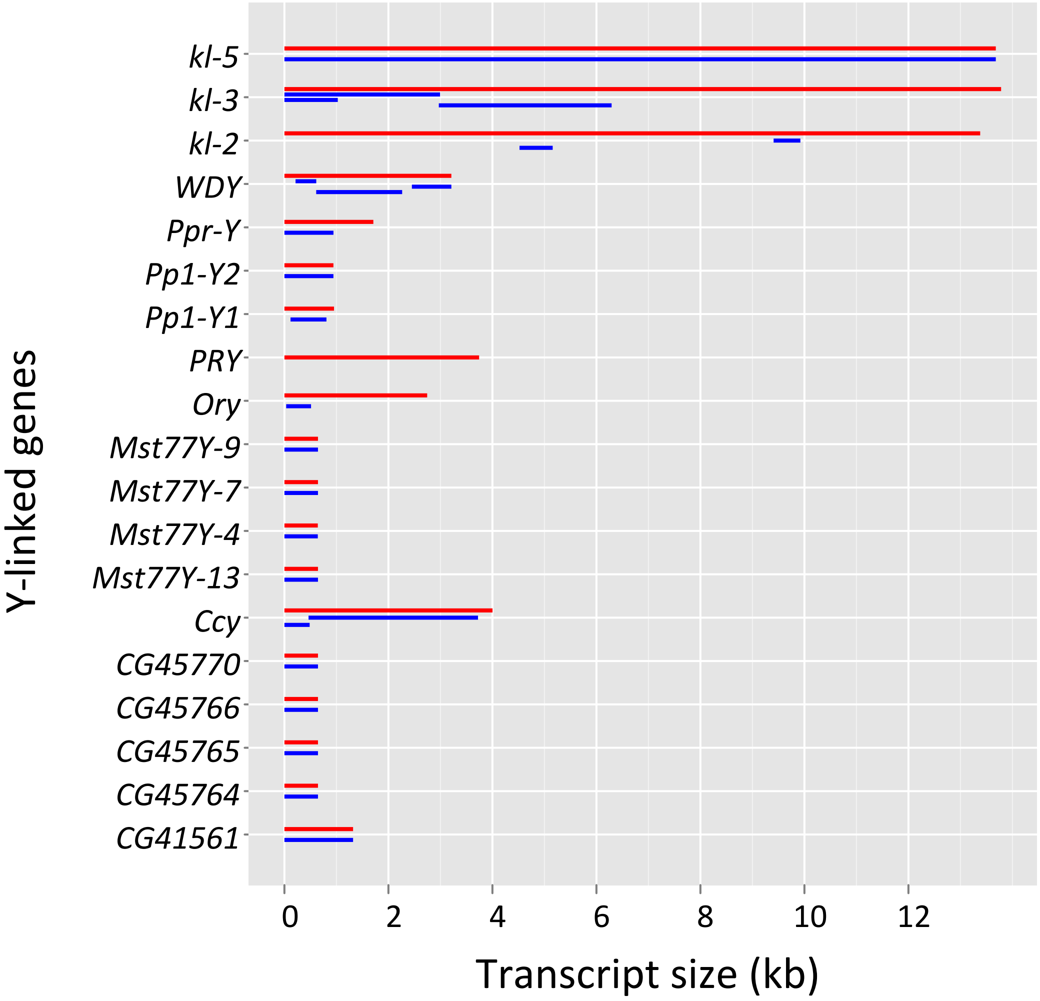
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	uncharacterized protein Dmel_CG41561 [Drosophila melanogaster]	897	897	100%	0.0	100%	NP_001104078.2
☐	MIP14882p [Drosophila melanogaster]	483	483	53%	3e-169	100%	ADB91442.1
☐	uncharacterized protein Dere_GG26511 [Drosophila erecta]	305	305	56%	2e-98	62%	XP_015008604.1
☐	uncharacterized protein Dere_GG27111 [Drosophila erecta]	193	193	43%	1e-55	58%	XP_015008591.1
☐	uncharacterized protein Dyak_GE28614 [Drosophila yakuba]	161	161	38%	2e-44	53%	XP_015045276.1
☐	uncharacterized protein LOC110187192 [Drosophila serrata]	92.0	92.0	19%	4e-17	54%	XP_020812251.1
☐	uncharacterized protein Dana_GF17090 [Drosophila ananassae]	85.1	85.1	22%	5e-14	49%	XP_001953716.2

Supplementary Figure 3. Blastp results for *CG41561*.



Supplementary Figure 4. PCR confirmation of Y-linked transcripts in *Phortica variegata, Teleopsis dalmanni*, *Ephydra hians* and *Themira minor*.

Y-linked transcripts amplify with male genomic DNA, but not with female genomic DNA.



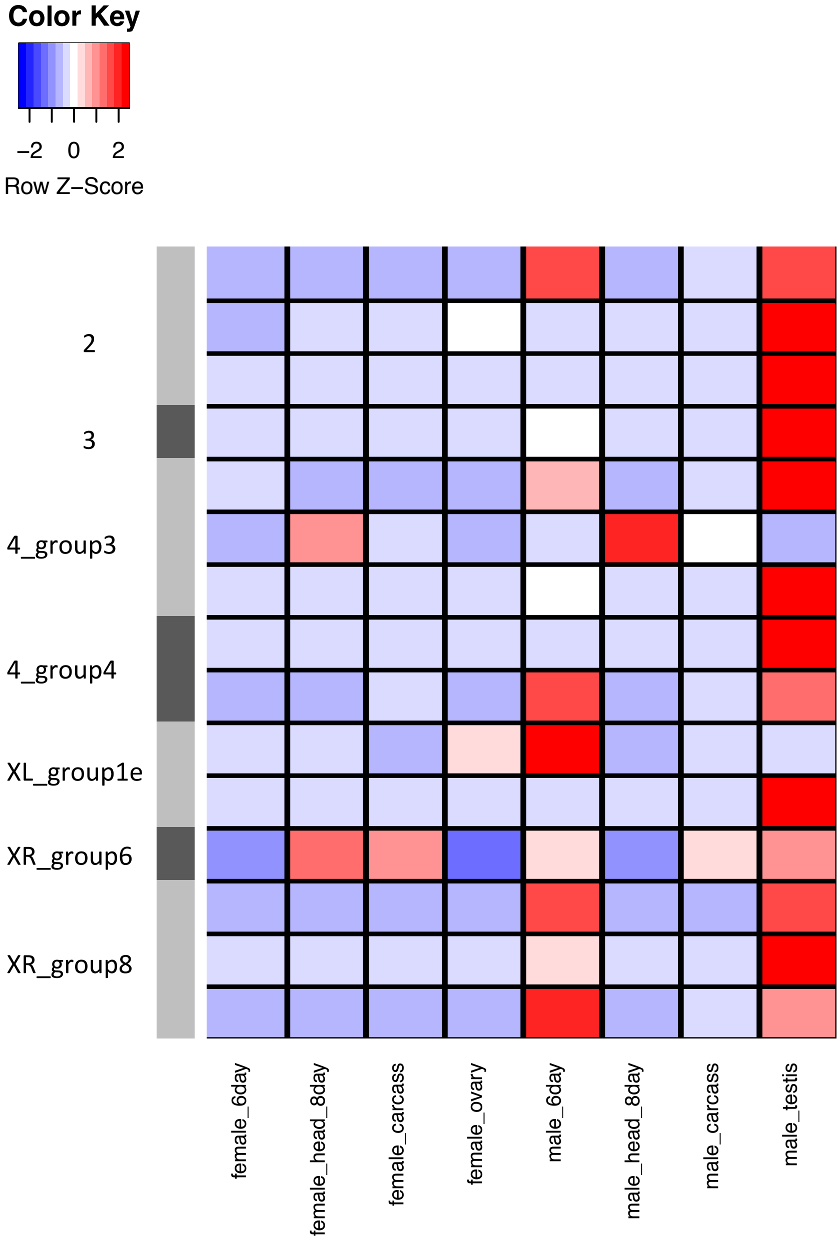
Supplementary Figure 5. Validation of our pipeline with subsampled *D. melanogaster* data.

Same as Supplementary Figure 1, but we subsampled our *D. melanogaster* data to match read counts with the species for which we have the lowest number of genomic read pairs (*Mayetiola destructor*).

```
Query: A.gambiae-TR578|c0_g1_i1 len=681 path=[659:0-680] [-1, 659, -2]
> X
dna:chromosome chromosome:AgamP4:X:1:24393108:1
Length = 24393108

Score = 632
Expect = 2e-161
Identity = 82.7648%
Strand = Plus/Minus
Aln Length = 553

Frame: 1 / -1

isTagged: , hitName: X
Query  129       TGATGGAGCGTGCGGAGAACATCACGCAGCGCAGGAAGCGTAGCTGGCTCCGGTTTCGTT  188
                 ||| || ||||||| ||  || |||||| | |||||||| |||||| ||||||||||| |
Sbjct  10685285  TGACGGTGCGTGCGCAGGGCAACACGCACCACAGGAAGCATAGCTGACTCCGGTTTCGAT  10685226


Query  189       TGATAGCGGCACAAATACTGTTCACCGCCACCCACCTTGGTGTTCTGCACGGCATCTCGA  248
                 |||||||||||||||||||||| | ||||||||   |||| |||||   |||||||||||
Sbjct  10685225  TGATAGCGGCACAAATACTGTTAATCGCCACCCTGTTTGGCGTTCTTATCGGCATCTCGA  10685166


Query  249       TCTTGAGGTATTTGAGAGCATGCAGGTCTGCGAAGGCCGGCCAATTGACGGCCACCGAAC  308
                 ||||||||||||| ||||| |||                  |||||||| ||||||||||
Sbjct  10685165  TCTTGAGGTATTTAAGAGCCTGC------------------CAATTGACCGCCACCGAAC  10685088


Query  309       CTACCGTAAA-GTTTTCCAGCCGCGGCAGCTTAATGTTCGCGAACGAAAGGGGCGTAGTG  367
                 | ||||| |  |||||||||||||||||  | ||||| |||||||||| ||| ||| |||
Sbjct  10685087  CCACCGTCAGCGTTTTCCAGCCGCGGCAATTCAATGTCCGCGAACGAATGGGACGTGGTG  10685028


Query  368       CTGTACAAAATTTTCAACTGGCGCAAGCCGGGCGGGTTCGCTTACAAGCGCAATGGGTTG  427
                 |||  ||||||||||| |||||||||||||| ||| ||||||  |||||||||||||||
Sbjct  10685027  CTGAGCAAAATTTTCAGCTGGCGCAAGCCGGACGGATTCGCTAGCAAGCGCAATGGGTTT  10684968


Query  428       CCGCTGCACTCTGCCACGTTAAATATGTACAAGTCCTCCAGCAGGACACACGATTGCCCG  487
                 ||||||||||||||||||||||| |||| ||||||||||||| |||||||||||| ||||
Sbjct  10684967  CCGCTGCACTCTGCCACGTTAAAGATGTGCAAGTCCTCCAGCTGGACACACGATTTCCCG  10684908


Query  488       ATGGCCTGCAGTACG---TCCAACACGGTATAAAATTGCAGGCGCAGCTTCTTTGCTGCT  544
                 |||||||||||||||   ||| || |||||||||||||||||||||||||||||   |||
Sbjct  10684907  ATGGCCTGCAGTACGCTTTCCGACGCGGTATAAAATTGCAGGCGCAGCTTCTTT--AGCT  10684846


Query  545       GACTCAACTGCCGGGaaaaaaaCAGTACCTTGTGCAGTACCTTGCTCAACTGTAGGTGTT  604
                   ||||||||||||   |||||||||||||| ||  |||| |    ||||    ||||||
Sbjct  10684845  AGCTCAACTGCCGG---AAAAACAGTACCTTTTGTTGTACAT---CCAAC----GGTGTT  10684776


Query  605       TTAGTGAACGTTGGCTTCAACTCGACCTCTATTTCGTGGCTTATGATCAGCTGCTCGAGC  664
                   ||||||||||||| ||||||| |||||  |||||||||||||||||||||||| ||||
Sbjct  10684775  GGAGTGAACGTTGGCATCAACTCAACCTCCGTTTCGTGGCTTATGATCAGCTGCTGGAGC  10684716


Query  665       TGCTTCAGCTGCCAGTA                                            681
                 |||||||||||||||||
Sbjct  10684715  TGCTTCAGCTGCCAGTA                                            10684699
```

Supplementary Figure 6. Alignment between the putative Y-linked transcript *A. gambiae* TR578 and an X-linked region in *A. gambiae.*

The transcript TR578 is homologous to parts of the previously described *YG1*/*YG2* genes.



Supplementary Figure 7. Expression patterns of autosomal or X-linked paralogs of testis-expressed Y-linked transcripts in *D. pseudoobscura.*

Most testis-expressed genes that were acquired or maintained on the neo-Y of *D. pseudoobscura* ancestrally have testis-biased expression (as inferred based on expression patterns of their paralogs). Each row corresponds to a gene, and genes are grouped by chromosomal location in *D. pseudoobscura.*

Supplementary Tables

Supplementary Table 1. Overview of all datasets used for this study, including Genbank accession numbers.

species	genomic data		transcriptome data	
	male	female	male	female
Chironomus riparius	SRR1738174	SRR1738173	PRJNA385725*	PRJNA385725*
Chaoborus trivittatus	SRR1738213	SRR1738278	PRJNA385725*	PRJNA385725*
Anopheles gambiae	SRR1509742	SRR1508169	SRR535750	SRR953402
Aedes aegypti	SRR1738168	SRR1738167	SRR924021 (testis)	SRR1585315
Clogmia albipunctata	SRR1738153	SRR1738152	PRJNA385725*	PRJNA385725*
Tipula oleracea	SRR1738202	SRR1738201	PRJNA385725*	PRJNA385725*
Coboldia fuscipes	SRR1738157	SRR1738156	PRJNA385725*	PRJNA385725*
Mayetiola destructor	SRR1738190	SRR1738189	SRR1738673	SRR1738672
Condylostylus patibulatus	SRR1738159	SRR1738158	PRJNA385725*	PRJNA385725*
Megaselia abdita	SRR1738192	SRR1738191	PRJNA385725*	PRJNA385725*
Themira minor	SRR1700645, SRR1700634	SRR1700632, SRR1700633	SRR1700682	SRR1700646
Bactrocera oleae	SRR826808	SRR826807	PRJNA385725*	PRJNA385725*
Teleopsis dalmanni	SRR1738200	SRR1738199	SRR1738676	SRR1738677
Liriomyza trifoli	SRR1700531	SRR1700530	SRR1700443	SRR1699519
Ephydra hians	SRR1738182, SRR1738176	SRR1738181, SRR1738175	SRR1738666	SRR1738664
Phortica variegata	SRR826813	SRR826812	SRR1738675	SRR1738674
Drosophila pseudoobscura	SRR1738164	PRJNA385727*	SRR357403	SRR357405
Drosophila miranda	SRR1738163	SRR1738162	SRR364798	SRR364800
Drosophila melanogaster	SRR1738161	SRR1738160	SRR1197415	SRR1197317
Drosophila busckii	SRR826814	SRR826809	SRR1804796	SRR1805120
Drosophila albomicans	SRR1738314	SRR1738289	SRR402049	SRR402050
Sarcophaga bullata	SRR826794	SRR826793	PRJNA385725*	PRJNA385725*

* These data were newly collected for this study

Supplementary Table 2. Accession numbers of male and female genomic *D. melanogaster* reads used to study the newly identified Y-linked gene *CG41561*. Male and female strains were chosen at random from the NCBI SRA database.

SRA Accession	Sex	Strain	Source	Coverage (% of transcript)
ERR701706	Male	Iso1	Bloomington	Yes (75%)
ERR701712	Male	nos-GAL4; UAS-DCR2	Greg Hannon Lab	Yes (60%)
SRR1525699	Male	Wild caught	USA: Bowdoinham, ME, Paul Schmidt	Yes (84%)
SRR1525770	Male	Wild caught	USA: Linvilla, PA, Paul Schmidt	Yes (90%)
SRR1738161	Male	Canton-S Lab strain	Stock center	Yes (80%)
ERR705952	Female	Haddrill France 31	Isofemale line collected in Montpellier, FRANCE	None
ERR705984	Female	Haddrill Georgia Pool15	15 isofemale strains collected in Athens, Georgia	None
SRR2134629	Female	Canton-S	Mark Biggin and Eisen Lab, UC Berkeley	None
SRR492060	Female	Z30	Carnegie Mellon University, Joel McManus	None
SRR1738160	Female	Canton-S Lab strain	Stock center	None

Supplementary Table 3.

Genome and transcriptome assembly statistics for Diptera species considered, and false-positive rate of our bioinformatics pipeline (as measured by the number of male- and female-specific transcripts for each taxon). Species are ordered as in ref. [1].

Species	Sex chromosome	Genome statistics					Male-specific Transcriptome statistics				Sex-specific contigs	
		N50 (bp)*	Total assembled genome size (Mb)	Number of raw sequencing reads pairs (x 10^6)	coverage	% reads mapping to assembled genome	Transcript N50 (bp)	median transcript length	average transcript length	Total size male-specific transcriptome	# male-specific contigs	# female-specific contigs
Chironomus riparius	homo	14,912	185.4	12.7	~12X	67.5	350	299	364.55	4,107,046	8	6
Chaoborus trivittatus	homo	2,743	408.9	31.5	~14X	47.3	320	287	334.1	6,158,470	17	2
Anopheles gambiae	XY	7,484	219.5	15.6	~13X	62.1	401	318	399.95	6,889,454	3	0
Aedes aegypti	homo	1,619	726.5	27.3	~7X	35.8	438	340	428.96	8,617,313	2	11
Clogmia albipunctata	homo	11,179	310.8	19.0	~11X	60.2	392	305	411.9	5,987,392	21	0
Tipula oleracea	XY	1,603	534.8	18.7	~6X	35.8	473	342	444.94	17,549,300	4	2
Coboldia fuscipes	XY	241,756	102.8	17.7	~31X	88.6	431	333	438.98	82,089	0	0
Mayetiola destructor	X0	14,122	154.5	7.7	~9X	69.5	349	302	356.28	7,117,333	0	0
Condylostylus patibulatus	XY	1,571	569.4	53.0	~17X	38.8	360	302	369.35	5,251,364	8	20
Megaselia abdita	homo	4,378	485.8	38.8	~14X	43.7	387	306	404.29	3,790,178	0	0
Themira minor	XY	2,212	114.9	16.9	~27X	67.8	355	301	366.09	3,861,476	16	1
Bactrocera oleae	XY	8,832	390.6	40.0	~18X	90.2	339	293	354.72	2,093,914	2	1
Teleopsis dalmanni	XY	4,268	575.0	39.2	~12X	69.9	470	341	449.72	8,185,400	22	2
Liriomyza trifolii	XY	1,633	125.9	19.3	~28X	45.5	363	307	370.54	10,049,396	1	0
Ephydra hians	XY	2,261	487.2	32.4	~12X	71.3	394	316	396.62	10,619,382	9	2
Phortica variegata	XY	37,343	253.2	13.0	~9X	54.8	1148	380	707.35	32,025,171	16	3
Drosophila pseudoobscura	XY	55,872	134.4	15.7	~21X	76.1	603	368	526.76	39,006,511	76	3
Drosophila miranda	XY	24,439	175.1	8.2	~8.5X	62.5	684	370	565.95	53,483,655	433	0
Drosophila melanogaster	XY	88,921	120.2	15.1	~23X	81.1	530	316	496.27	10,464,410	29	3
Drosophila busckii	XY	36,315	121.1	25.8	~38X	78.3	445	327	434.28	14,172,673	139	0
Drosophila albomicans	XY	35,303	158.2	12.0	~14X	68.4	348	298	360.53	6,173,786	61	0
Sarcophaga bullata	XY	1,828	484.6	38.0	~14X	66.4	381	298	388.54	1,242,166	0	4

*female-only genome assembly

Supplementary Table 4. PCR primers amplifying male-specific products in Diptera species (i.e. Y-linked transcripts).

species	transcript	for_primer	rev_primer	PCR product	
				(bp)	notes
Themira minor	CL166Contig1	TCGACGTTGTGCTCTTTGAG	GAGCCACGTGAATGTTCAGA	225	protein-coding
Themira minor	TR7942	CAGTCGGATGACTTCGTTCC	TTTGAATTTTCGTCGTGTGC	186	
Themira minor	TR242	TGTCCTCCTTTGGGTTTCAC	ACCGCAAGTTTCTCGGAATA	175	
Themira minor	TR3320	AGAGCGTCCTTCCTTTGTGA	CGATGACGGTGATCTTGTTG	195	
Themira minor	CL55Contig1	AGTTTCGGACAAGAGCAGGA	AATATCCGTTTGGTGCTTGC	199	
Themira minor	CL11Contig2	GGCCTTCGGATTTTAGGAAG	AAGACTACTTGCGCGACGAT	186	
Teleopsis dalmanni	TR12456	TTTGGCGTTATGTTCCTGGT	CCGCTATTTTCCCCCATAGT	237	protein-coding
Teleopsis dalmanni	TR8004	CATCAAAGAGTGGCACAGGA	AACCTTTTGCTGCGCTTCTA	164	protein-coding
Teleopsis dalmanni	TR2473	TTGCCTGGAAAAAGAATTGG	TGTTAGCTTCATGCGAAACG	182	protein-coding
Teleopsis dalmanni	TR4930	ACAGGCTCGCTAAGTTGGAA	ACCCTAACGGATCCACACAA	225	protein-coding
Teleopsis dalmanni	CL466Contig1	CGATTACCGATTGCTCCATT	AGGCACTACCGATAGCGAGA	192	protein-coding
Teleopsis dalmanni	TR6579	CGTATGTCTCGCCGAGTGTA	ACAAGCGCTTCAGATCTTCC	151	protein-coding
Teleopsis dalmanni	TR2554	TTTACAGTTCCCTCGGATGC	TGCATCATTTGAAAGGGATTT	195	protein-coding
Teleopsis dalmanni	CL291Contig1	TCGGGAATAACGCAGGATAC	CTTGGGGCATCATTTTGTTT	218	protein-coding
Teleopsis dalmanni	TR6646	TCACCAATGCACCAACATTC	CGTTGCAAATAAGCATCCAA	209	
Teleopsis dalmanni	TR8935	TCCATTTGGAGTGGACCTGT	GCACATGCTTCGAATTGTTG	154	
Ephydra hians	CL425Contig1	CGTACCAGAGCAGACACCAA	AACCAAACAAACAGAGCTTGC	195	
Ephydra hians	CL517Contig1	TGAAGAACTTTGTTGGTGCATT	CCGTCCGGAATGTGTTTATG	249	
Ephydra hians	CL517Contig1*	ACGGGAGGAAATTCGAGTAA	TTGGTGTCTGCTCTGGTACG	246	
Ephydra hians	CL624Contig1	AAATAAAGCCGTTGAGCAGT	AACCTTTGCAAAACGCTATC		
Ephydra hians	CL624Contig1*	ATTATTTCCGTCGCACTTTC	TTTGAGCACAGTCACCAAAC		
Ephydra hians	TR1001	CTTTGATCTTGGCCGTGTTT	AAGAACTTTTCTGATGCATTGCT	213	
Ephydra hians	TR10871	GCATCAGTAAAAGGGGCAAA	GACGCCTAAGGCCATCATTA	160	protein-coding
Ephydra hians	TR18873	CAAAGGGTGCGTCCATTATC	ACAGCGTTTTGTCTCCGTTT	186	
Phortica variegata	CL1175Contig1	CCTATTGCAGCTGATGACCA	TGGCACAAGTTTCAGCAGAC	295	protein coding
Phortica variegata	CL542Contig1	GTCGCCAATGTGACTCTGAA	CTGCTGTTTTCAGCCATCAA	299	protein coding
Phortica variegata	TR15734	TCTCCCTGAATTTACCAAAGGA	TTTGCGGTTCGGAAAAATTA	293	
Phortica variegata	TR18029	TTTAACAAATTCGCGGTCAT	TGCATTATAAGATGGCAGGA	97	
Phortica variegata	TR18029*	TTACAGCGGTACAACCAAAA	GACAACTCAGATTCGGACATT	101	
Phortica variegata	TR25433	TTGGGTCATGCATAGAGAAA	AGAACTTGTGGGAGAATGGA	100	
Phortica variegata	TR25433*	ATCCAAGACTGGAGGTTCAA	GCCAGATCGCATTTCTCTAT	81	
Phortica variegata	TR25932	CGTCTCCTCCTCATCTTTGC	TCACCGTTTCTCACAGAGCA	278	
Phortica variegata	TR28140	ATTCGTTCAGGCTCATTTGT	GCCACGTTTGATCCAATACT	77	
Phortica variegata	TR28140*	GCACAGAAAAACCGGAGTAT	TGTTTTGCCTTCTATGTGCT	83	
Phortica variegata	TR6288	AAATGTGGGAGAATTGGAAA	GAACTAAACTGGGGAGCTGA	299	
Phortica variegata	TR6288*	TATTCCAACCCCTGGTAAAA	AAGGCGTGTATTTCCTTCTG	307	

Supplementary Table 5. Accession numbers of tissue-specific expression data.

species	Somatic tissues		Gonads	
	Male head	Female head	Testis	Ovary
Themira minor	SRR1700709	SRR1700675	SRR1700693	SRR1700647
Teleopsis dalmanni	SRR1184533	SRR1184534	SRR1184544	SRR1184546
Ephydra hians	SRR1738668	SRR1738667	SRR1738671	SRR1738669
Drosophila pseudoobscura	SRP001791	SRP001791	SRR357404	SRR357400
Drosophila miranda	SRR4416188	SRR4416186	SRR364799	SRR364801
Drosophila melanogaster	SRR070400, SRR070416	SRR070430, SRR100282	SRR100276, SRR070422	SRR070396, SRR070417
Drosophila albomicans	SRR4416171	SRR4416190	SRR4416172	SRR4416191

Chapter 3

Assembly of a young Drosophila Y chromosome using Single-Molecule sequencing and Chromatin Conformation capture

Shivani Mahajan, Kevin Wei, Matt Nalley, Emily Brown, Doris Bachtrog

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

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 repeat-rich Y chromosomes are typically ignored from sequencing projects. Here, we assemble the genome of *Drosophila miranda* using long reads for contig formation, short reads for consensus validation, and scaffolding by optical and chromatin interaction mapping and BAC clone sequencing. Our assembly covers large fractions of repetitive DNA, including >100Mb of the recently formed neo-Y chromosome. The neo-Y chromosome, which diverged from its homolog, the neo-X, only about 1.5 million years ago, has dramatically increased in size by almost 3-fold, due to the accumulation of repetitive sequences.

Introduction

Sex chromosomes are derived from ordinary autosomes, yet old X and Y chromosomes contain a vastly different gene repertoire (Bachtrog *et al.* 2014). In particular, X chromosomes resemble the autosome from which they were derived, with only few changes to their gene content (Vicoso and Charlesworth 2006). 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 (Bachtrog 2013); old Y chromosomes typically contain only few genes but vast amounts of repeats.

The decrease in sequencing cost and increased sophistication of assembly algorithms for short-read platforms has 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 (Hoskins *et al.* 2002; Schatz *et al.* 2010;

Khost *et al.* 2017). 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 (Soh *et al.* 2014), or gene conversion within palindromes retarding Y degeneration in primates (Skaletsky *et al.* 2003). However, the repeat-rich nature of Y chromosomes has hampered their evolutionary studies in most organisms.

Here we present a near-finished reference genome for the recently formed neo-Y chromosome of *Drosophila miranda* using a combination of long-read single-molecule sequencing, high-fidelity short-read sequencing, optical mapping, 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, due to its recently evolved neo-sex chromosome system (see **Figure 1**). In particular, chromosomal fusions within *D. miranda* have resulted in the recent sex-linkage of former autosomes, with chromosome XR becoming sex-linked about 15 million years ago, and the neo-X and neo-Y becoming sex chromosomes only about 1.5 million years ago (Bachtrog and Charlesworth 2002). These former autosomes are in the process of evolving the stereotypical properties of ancestral sex chromosomes (Zhou *et al.* 2013; Ellison and Bachtrog 2013), and allow the investigation of the functional and evolutionary changes occurring on differentiating sex chromosomes.

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 of 5,007 bp and a total assembled genome size of 112 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 (about 50% of its DNA is derived from repeats) posed a particular challenge to its assembly using short reads, and the neo-Y assembly was particularly highly fragmented and incomplete, consisting of 36,282 scaffolds, and a scaffold N50 of only 715 bp (Zhou and Bachtrog 2012). Here, we assemble the genome of *D. miranda* using long reads for contig formation, short reads for consensus validation, and scaffolding by optical and chromatin interaction mapping and BAC clone sequencing. Our assembly covers large fractions of repetitive DNA, including >100Mb 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 Y chromosome evolution in this species.

Results

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), shot-gun BAC clones sequencing (Illumina HiSeq) and Hi-C (**Supplementary Table S1**).

Assembly of these complementary data types proceeded in a stepwise fashion (**Figure 2**), similar to a recent approach (Bickhart *et al.* 2017), to produce progressively improved assemblies (**Table 1**). Briefly, we produced two initial assemblies of the PacBio data alone using the *Falcon* (Chin *et al.* 2016) and *Canu* (Koren *et al.* 2017) assembler, and double-merged the resulting assemblies with *Quickmerge* (Chakraborty *et al.* 2016). 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 (to avoid cross-mapping of short read Hi-C data, see **Figure 3**), and clustered into chromosome-scale scaffolds using Hi-C data (**Table 1**, **Figure 2**). Mapping of Illumina reads also allowed us to identify and remove contigs that resulted from uncollapsed haplotypes (**Supplementary Figures S1, S2).** X-linked and autosomal contigs were scaffolded with female Hi-C libraries (**Figure 4, 5**), while Y-linked contigs were clustered using male Hi-C libraries (**Figure 4, 5**). Visual inspection of contact probability maps allowed us to identify a few mis-assemblies, which were manually corrected followed by re-scaffolding (see **Supplementary Table S2**). To assess quality, the resulting assembly was validated using short read Illumina mapping, comparison to optical mapping data (**Supplementary Table 1** & **Figure 6**), and sequenced BAC clones from the MSH22 strain (**Supplementary Table 1, Figure 7**) and previous assemblies (*D.miranda*v1 (Zhou and Bachtrog 2012) (**Figure 8** and **Supplementary Figures 3a-3f);** *D. pseudoobscura (***Figure 9** and **Supplementary Figures 4a-4e**) and *in situ* hybridization data for *D. miranda* (Bartolomé and Charlesworth 2006).

To maximize accuracy of the final reference assembly, errors were manually curated before final gap filling and polishing. Our final assembly, D.mir2, totaled 287Mb of sequence with a scaffold NG50 of 35.3 Mb (**Table 1**). D.mir2 comprises just 103 scaffolds and 120 gaps. We used two approaches, REPdenovo (Chu *et al.* 2016) and RepeatModeler (Smith and Hubley) to annotate repeats in the *D. miranda* genome, and Maker (Campbell *et al.* 2014) to annotate genes.

Assembly benchmarking and comparison to reference

Mapping to previous *D. miranda* assembly: The D.mir1.0 reference assembly was generated from paired-end short reads using the SOAPdenovo assembler, and cross-species scaffold alignments to the *D. pseudoobscura* (Zhou and Bachtrog 2012). Paired-end read sequences used to create the D.mir1.0 reference assembly 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 general improvement over D.mir1.0 (**Table 3**), with fewer putative translocations (36 versus 17,764), deletions (229 versus 6075) and duplications (8 versus 1703).

The initial *D. miranda* genome was scaffolded *using D. pseudoobscura*, and genome-wide alignments between our current *D. miranda* assembly and D.mir1.0 reveals dozens of inversions (**Figure 8**; **Supplementary Figures 3a-3f**).

Mapping to sequenced BAC clones and optical mapping. We assessed large-scale structural continuity of each assembly by aligning BAC clone sequences and identifying structural variants and potential mis-assemblies. In total, we shotgun sequenced 383 BAC clones, which should cover roughly 1/4 of the *D. miranda* genome. **Figure 7** shows mapping of BAC clones to the assembly. The majority of BAC clones maps contiguously across the genome, to a unique position, supporting that our genome assembly is of high quality. Similarly, most of our genome is covered by optical mapping data (**Figure 6**).

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

The presence of its recently formed neo-sex chromosomes have established *D. miranda* as an important model system, yet the assembly of both the neo-X and neo-Y proved particularly challenging to short read technology. On one hand, the high level of sequence identity at homologous regions (98.5%) implied that many reads could not be unambiguously assigned to either the neo-X or neo-Y chromosome. On the other hand, the drastic accumulation of repeats on the neo-Y resulted in a highly fragmented and incomplete assembly of the neo-Y (**Table 2**). In particular, the original neo-Y assembly consisted of 36,282 scaffolds, and a scaffold N50 of only 715 bp (Zhou and Bachtrog 2012). In our current assembly, most of the Y chromosome is contained in only two large scaffolds (54.2 Mb and 36.6 Mb).

Assembly of highly repeat-rich regions

In addition to recovering the repeat-rich, recently formed neo-Y chromosome, our assembly also contains large blocks of pericentromeric DNA. In particular, we assemble 41 Mb of pericentromeric repeats and telomeres (**Table 4**). In contrast, the previous assembly based on only Illumina reads recovered less than 1Mb of peri-centromeric DNA.

Repeat & gene content of *D. miranda*

We generated a *de novo* repeat library to annotate repeats. We used two approaches, REPdenovo (Chu *et al.* 2016) and RepeatModeler (Smith and Hubley) to annotate repeats in the *D. miranda* genome, and Maker (Campbell *et al.* 2014) to annotate genes. **Figure 10** shows an overview of genes and repeats annotated across the *D. miranda* genome. We identified a total of 17,745 genes and 43.7% of the genome was annotated as repeats.

Identifying orthologs between *D. pseudoobscura* and *D. miranda* proteins

We identified orthologs 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.

Contrasting gene content on neo-X and neo-Y

Y chromosome evolution is characterized by massive gene loss (Bachtrog 2013). To identify genes that have been lost from the neo-Y chromosome, we used BLAST to identify annotated genes that are present on the neo-X in *D. miranda* and in *D. pseudoobscura*, but completely absent from the neo-Y. 61 out of the 79 genes identified as missing from the neo-Y but present on the neo-X are on the homologous chromosome (Muller C) in *D. pseudoobscura,* suggesting that they were ancestrally located on the neo-sex chromosomes but lost from the neo-Y. Most of the remaining genes were found to have duplicated onto the neo-X chromosome, but also retained a copy of their ancestral location shared with *D. pseudoobscura*.

We also identified one gene (FBgn0246393) that is present on the neo-Y and on Muller C in *D. pseudoobscura*, but has lost its neo-X copy in *D. miranda.* This gene has no annotated *D. melanogaster* ortholog but it is expressed in testes & imaginal disc in *D. pseudoobscura*, suggesting that it has a male-specific function.

We also identified genes that are present on the neo-X but absent on the neo-Y independent of using orthology to *D. pseudoobscura.* In particular, we annotated 2,373 genes on the neo-X chromosome and 4,801 genes on the neo-Y and Y contigs (counting paralogs as separate genes for both chromosomes). We aligned the neo-X transcripts to the Y/neo-Y transcripts using blast (allowing partial alignments and minimum percent identity equal to 70%), and identified 244 annotated genes that are present on the neo-X (again counting paralogs as separate genes) but absent on the neo-Y.

Since Maker may fail to annotate certain genes (or some genes may have become pseudogenized and are not annotated by Maker), we also aligned neo-X genes directly to the neoY/Y-linked contigs in our assembly using blast (allowing partial alignments and minimum percent identity equal to 70%). We identified 106 genes present on the neo-X (counting paralogs as separate genes) that are absent from the neo-Y.

Gene content evolution on the Y/neo-Y

Of the 4,801 annotated genes on the Y/neo-Y scaffolds (including 3,689 on the two largest neo-Y scaffolds), 4,644 mapped to other annotated genes in the genome (allowing partial alignments and percent identity cutoff of 70%). 3,468 of these neo-Y genes mapped to annotated genes located on the neo-X (which has 2,373 total annotated genes, 106 of which do not align to Y/neo-Y scaffolds). This indicates that there has been widespread amplification of genes on the neo-Y chromosome. Indeed, using Illumina genomic reads, we were able to confirm that several genes have amplified on the Y chromosome.

We were unable to identify paralogs (as annotated by Maker) for 157 neo-Y/Y genes elsewhere in the genome. Again since Maker may fail to annotate some genes (or pseudogenes), we directly aligned the 4,801 Y/neo-Y-linked genes to the autosomes and XL, XR and neo-X genomic scaffolds. 4,790 of these Y/neo-Y mapped to autosomal/XL/XR/neoX-linked scaffolds (allowing partial alignments and percent identity cutoff of 70%), and only 11 genes did not have any paralogs in the genome. Of these, one gene (Fbgn0246875) was previously found to be located on the ancestral Y of *D. pseudoobscura* (Mahajan and Bachtrog 2017).

Divergence between single-copy genes on the neo-X and the neo-Y

For this analysis we ignored multicopy genes and only used loci, which had one annotated copy each on the neo-X and the neo-Y (1207 gene pairs in total). We extracted coding sequences for these genes using the gffread utility from Cufflinks (Trapnell *et al.* 2012), used Prank (Löytynoja 2014) to generate sequence alignments and KaKs_Calculator (Zhang *et al.* 2006) to calculate Ka, Ks and Ka/Ks values between homologous neo-X/neo-Y genes (**Table 5**). We also restricted our analysis to 310 gene pairs that aligned over 99.5% of their length and the alignment started with ATG (conserved set, **Table 5**). As expected, genes that align over most of their length show higher levels of sequence evolution and constraint.

Rates of Divergence between *D. pseudoobscura* and *D. miranda*

We calculated Ka,Ks and Ka/Ks values for each chromosome to estimate rates of divergence between genes in *D. pseudoobscura* and *D. miranda.* Based on our previous orthology calls, we extracted genes that had only one annotated copy on each chromosome and aligned them to the corresponding ortholog in *D. pseudoobscura* (from Flybase using only the longest CDS) with Prank and used KaKs_Calculator to estimate rates of protein evolution (**Figure 11**).

Conclusion

Our new assembly provides a highly improved *D. miranda* genome assembly. This assembly will provide the basis for further evolutionary and functional research on the newly formed neo-sex chromosomes of *D. miranda*.

Materials and Methods

Fly strain

We chose the inbred MSH22 strain for *D. miranda*, which was previously used to generate a BAC library (Bachtrog *et al.* 2008), and for genome assembly using short Illumina reads (Zhou and Bachtrog 2012).

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 Gbp of data with a read N50 of 17.1 Kbp

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). The IrysView (BioNano Genomics) software package was used to produce single-molecule maps and *de novo* assemble maps into a genome map (**Table 1**).

PacBio assembly

40x error corrected reads were used to build an initial PacBio assembly using the *Falcon* assembler (Chin *et al.* 2016). 28Gb long reads (NR50 = 17116 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) (Chin *et al.* 2016) running on Sun Grid Engine in parallel mode. For assembly, reads longer than 10Kb and 17Kb were used as seed reads for initial mapping and pre-assembly. 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 (Chin *et al.* 2013), followed by the software Pilon (Walker *et al.* 2014) 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* (Koren *et al.* 2017) with default parameters. This assembly consisted of 521 scaffolds totalling 296,012,170 bp, with an N50 value of 3,884,273 bp. These two assemblies were the merged using Quickmerge (Chakraborty *et al.* 2016), with default parameters. The resulting merged assembly was then merged a second time to the finished Falcon assembly, producing a Quickmerged 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 3rd instar larvae of MSH22. Briefly, chromatin was isolated from male and female 3rd 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 HpychIV and HindIII libraries, respectively. For males, 28.0 and 179.2 million pairs were produced.

Hi-C-based proximity guided assembly (PG)

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 (Durand *et al.* 2016) to align female Hi-C reads to the autosomal/X-linked scaffolds and also to align a subset of male HiC reads (that don't 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 (Dudchenko *et al.* 2017). Using a custom Perl script, we then scaffolded the Pac-Bio 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 single-end mode on default settings. The resulting sam files were then used to generate genome-wide Hi-C interaction matrix using the program Homer. For visualization, we plotted the interaction matrix as a heatmap in R, with demarcations of the Pac-Bio contigs and Hi-C scaffolds. Iteratively, we visually examine the heatmap to identify possible anomalies for scaffolding errors and manually curate the .asm file output to improve the heatmap.

Inferring Heterozygosity

We mapped paired-end illumina libraries from 5 individuals to the assemblies using bwa mem (Li and Durbin 2009) on default settings. We then use GATK (DePristo *et al.* 2011), following their best practices recommendations, to identify polymorphic sites. The resulting vcf file was parsed for heterozygous sites with genotype quality > 20. The numbers of these sites were then binned along the chromosome in 100kb windows.

BAC clone DNA isolation and sequencing

Bacteria were cultured in Terrific Broth with 25 µg/mL chloramphenicol. From glycerol stocks, 200 µl starter cultures in 2.0 mL deep-well plates were inoculated, covered with AeraSeal gas permeable films, incubated at 37ºC with shaking for 7-8 hours, then vortexed briefly to resuspend settled cells. Overnight cultures (500 µl) were inoculated with 0.5 µl from starter cultures, covered with AreaSeal films, and incubated at 37ºC with shaking for 12-14 hours. These cultures were centrifuged for 10 minutes to pellet cells. Media was poured from the deep-well plates, which were then tapped upside down on paper towels to completely drain the pellets.

Solution II [NaOH (200 mM) and SLS (1%)] was prepared fresh from stocks of NaOH (1 M) and SLS (20%), and Solution III [KOAc (5 M, pH 5)] was cooled on ice. To cell pellets, 60 µl Solution I [Tris-HCl (50 mM, pH 8) and EDTA (50 mM)] was added with pipet mixing to resuspend the cells. Solution II (120 µl) was added without pipet mixing; the plates were covered with aluminum seals and inverted gently four times. After five minutes at room temperature, the plates were spun briefly before removing aluminum seals. Solution III (270 µl) was added without pipet mixing. The plates were sealed, inverted gently four times, chilled on ice for 10 minutes, and centrifuged for 1 hour. To 120 µl isopropanol, 200 µl supernatant was added using wide-bore pipet tips. The plates were sealed, inverted gently three times, and centrifuged for 1 hour to pellet BAC DNA. Pellets were drained, rinsed with 200 µl 70% ethanol, sealed, centrifuged for 5 minutes, drained, rinsed with 200 µl 80% ethanol, sealed, centrifuged for 5 minutes, drained again, and dried at 37ºC for 10 minutes. Qiagen EB (50 µl) was added to each well to dissolve the BAC DNA at room temperature overnight.

BAC clone mapping

For each BAC, 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 (Langmead and Salzberg 2012)and SAMtools (Li *et al.* 2009). The remaining trimmed, filtered reads were mapped to our *D. miranda* assembly using bwa (Li and Durbin 2009). The BAC's location was determined by filtering regions of high coverage (at least 50X mean) and significant length (at least 20-kb). First, regions with average coverage of at least 50X were extracted, and any regions within 250-kb of each other were merged using BEDtools (Quinlan and Hall 2010). (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 mapped to the vector and its mate mapped to our assembly. Filtered reads were mapped to pTARBAC-2.1 using bwa (Li and Durbin 2009), 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 4000-bp of the vector's start position were considered "start" reads, and reads mapping within 4000-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 (Quinlan and Hall 2010), 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, putative BAC regions were masked using BEDtools (Quinlan and Hall 2010), and the reads were mapped back to this masked assembly then filtered and merged as described above. Regions of primary and secondary mapping were reviewed using IGV to show that little cross mapping occurs in our assembly; after masking and re-mapping, we found significant mapping to homologous regions of the its homologous neo-sex chromosome.

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 >100kb regions where roughly half of the reads map to another region in the two mappings (see **Supplementary Figures S1, S2**). For erroneous duplications and mis-scaffolded contigs in the Pac-Bio assembly identified, we used IGV to visualize the quality of Illumina reads mapping and to determine the precise coordinates to modify (**Supplementary Table 2**). 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 are then removed. Because, mis-scaffolded contigs are typically caused by misassembly around repetitive elements, we therefore also separate two contigs based on visual inspection of non-uniquely mapping reads.

Repeat Annotation and masking

For repeat masking the genome, we annotated repeats using REPdenovo (downloaded November 7, 2016; (Chu *et al.* 2016)) and RepeatModeler version 1.0.5 (Smith and Hubley). We ran REPdenovo on on raw sequencing reads using the parameters MINREPEATFREQ 3, RANGEASMFREQDEC 2, RANGEASMFREQGAP 0.8, KMIN 30, KMAX 50, KINC 10, KDFT 30, GENOMELENGTH 176000000, ASMNODELENGTHOFFSET -1, MINCONTIGLENGTH 100, ISDUPLICATEREPEATS 0.85, COVDIFFCUTOFF 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 (Gramates *et al.* 2017) 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 with an N50 of 634, totaling 964,435 base pairs. Our RepeatModeler repeat annotation had 1,009 repeats with an N50 of 715, 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 (Smith *et al.*).

Structural Variant Calling for Quality Control

For the previous published genome assembly and the various intermediate assemblies produced during the making of the current version, we estimated quality statistics using the variant caller LUMPY (Layer *et al.* 2014). To do this, we first aligned reads from two separate male Illumina libraries (with 626bp and 915bp insert sizes respectively) to our current assembly and its intermediates using SpeedSeq, which does a BWA-MEM alignment and produces the discordant and split reads bam files needed to run lumpyexpress. Since the published genome is a female only genome, we also characterized structural vairants for it using only female illumina reads (three libraries with 285bp, 640bp and 918bp respectively)

We then ran the software lumpyexpress using the output bam files produced by Speedseq as input 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.

Gene annotation using Maker

To run Maker (Campbell *et al.* 2014), we first build a transcriptome assembly. 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 3rd instar larvae) were aligned to the genome assembly using HiSat2 (Kim *et al.* 2015) 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 (Pertea *et al.* 2015) 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 (Campbell *et al.* 2014) 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 (Korf 2004) and Augustus (Stanke and Waack 2003) 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 3rd 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 genes. We assigned putative functions to the genes annotated by Maker by first aligning them to the curated Uniprot protein sequences using blast (parameters evalue 1e-6 -seg yes -soft_masking true -lcase_masking -max_hsps 1 -num_alignments 1), followed by running the maker_functional_gff script. We were able to assign putative functions to 11,797 *de novo* annotated genes.

Identifying *D. pseudoobscura* orthologs for *de novo D. miranda* proteins

We identified orthologs by aligning *D. pseudoobscura* proteins to our list of *de novo* annotated *D. miranda* proteins. To do this, we first aligned *D. miranda* proteins (as query) to *D. pseudoobscura* proteins (as database) using BLAST using the following parameters allowing for multiple hits per *D.miranda* protein: blastp -db *Dpseudoobscuradatabase* -query *Dmirandaproteins.fa* -out blastoutput.txt -evalue 0.0000001 -outfmt 6 -seg yes -soft_masking true -lcase_masking -num_descriptions 50 -num_threads 24 -num_alignments 50**.** We then removed any hits that aligned with a percent identity less than 59%.

For each *D. miranda* protein, we identified the best hit by first sorting for the largest bit score, followed by sorting for the smallest e-value, followed by sorting for the highest percent identity, followed by sorting for the longest *D. pseudoobscura* protein, followed by name sorting the *D. pseudoobscura* protein.

We also aligned *D. pseudoobscura* proteins to *D. miranda* proteins using BLAT (parameters minScore=50 and minIdentity=60). We combined the bestHits with blast with the BLAT alignment results to get a final list of orthologs between *D. pseudoobscura* and *D. miranda.*

For 16,378 of the total 17,745 genes in our annotation we were able to reliably identify orthologs in the *D. pseudoobscura* annotation, and 1,367 genes were not previously annotated in the *D. pseudoobscura* genome. We used blastp to align protein sequences of these 1,367 genes to previously annotated *D. melanogaster* proteins. Keeping only blast hits with a minimum percent identity of 40%, we were able to identify *D. melanogaster* orthologs for 285 of these 1367 genes.

Therefore, for 1,082 genes we were unable to identify orthologs in neither the *D. pseudoobscura* nor the *D. melanogaster* genome.

Tables

Table 1 – Assembly statistics (including intermediate assemblies produced at different steps, the old published assembly and the current genome assembly).

Assembly	Contigs + Scaffolds	Scaffolds	Unplaced Contigs	Contig N50 (Mb)	Scaffold N50 (Mb)	Assembly Size (Mb)	Assembly in Scaffolds (%)
PacBio Falcon	625	NA	625	2242328	NA	273479696	NA
PacBio Canu	521	NA	521	3884273	NA	296012170	NA
Quickmerged	271	NA	271	5177776	NA	295243618	NA
PacBio+Hi-C	103	14	89	NA	37186217	288869284	96.49
D.mir_Zhou 2012 (artificially stitched)	536	6	530	NA	28826359	139692925	97.88
D.mir Zhou 2012	4766	NA	530	NA	0.15	139692925	NA
D.mir_current	103	14	89	NA	35263102	287222512	96.57

Table 2 – Assembly statistics of neo-Y and Y-linked scaffolds.

Assembly	Contigs	Scaffolds	Unplaced Contigs	Contig N50 (Mb)	Scaffold N50 (Mb)	Assembly Size (Mb)	Assembly in Scaffolds (%)
Quickmerged	193	NA	NA	2049273	NA	112862549	NA
PacBio+Hi-C	62	8	54	37186217	NA	112190153	95.52
D.mirZhou 2012	NA	36282	NA	NA	715	22000000	NA
D.mircurrent	62	8	54	NA	366373738	110513651	95.69

Table 3 – Structural variant identified using Lumpyexpress, by mapping two male MSH22 Illumina libraries back to the reference genomes (626bp and 915bp insert sizes).

Assembly	BND	DEL	DUP	INV	Errors per Mb
Canu	108	206	10	2	1.10
Polished Falcon	224	229	35	10	1.82
Quickmerged	64	348	17	2	1.46
Pacbio + HiC	62	292	17	2	1.29
Old D.miranda (Chromosomes + unplaced scaffolds)	14248	5780	1430	77	154.16
Old D.miranda (Chromosomes only)	17764	6075	1703	117	187.66
D.mir current	36	229	8	4	0.96

*BND = trans-contig association, DEL = deletion, DUP = duplication, INV = inversion

Table 4 – Comparison of current and previous assembly of *D. miranda.*

	D. mir_current				D.mirZhou 2012			
	Size in Mb	No. Scaffolds	Repeat %	Peri-centromere in Mb	Size in Mb	No. of Scaffolds	Repeat %	Peri-centromere in Mb
Total Genome	287.2	103	43.72	40.65	158.7	4766	7.27*	0.5
XL	25.3	1	18.27	0.55	22.1	1463	7.15	~0
XR	52.4	1	38.7	20	30.1	784	5.29	~0
Chr2	35.3	1	12.65	2	33.0	947	5.84	~0
Chr4	32.5	1	15.66	3.7	28.8	834	4.8	~0
dot	2.4	1	48.69	1.2	1.8	238	22.06	0.5
Neo-X	25.3	1	20.86	4	20.9	744	6.13	~0
Neo-Y	90.8	2	71.90	9.2	22	36,282	NA	NA

Table 5 – Rates of protein evolution between homologous gene pairs on the neoX-sex genes.

	All genes		Conserved genes	
	median	[min-max]	median	[min-max]
Ka	0.011	[0.0007-0.870]	0.008	[0.0009-0.067]
Ks	0.036	[0.004-3.551]	0.032	[0.003-0.485]
Ka/Ks	0.302	[0-4.191]	0.243	[0-3.186]

Figures

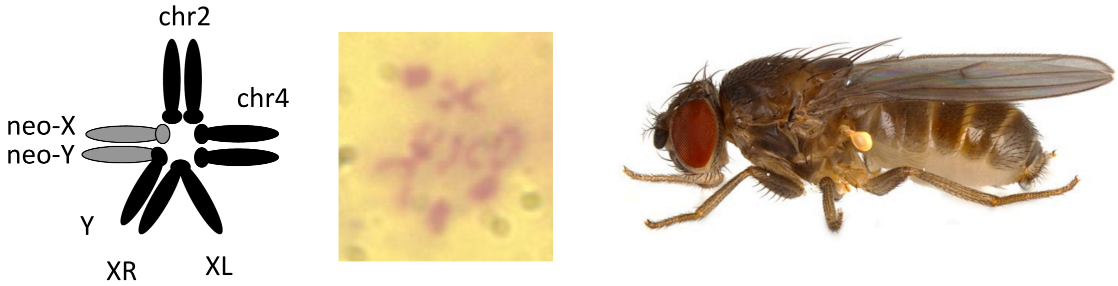


Figure1. Karyotype of *D.miranda.*

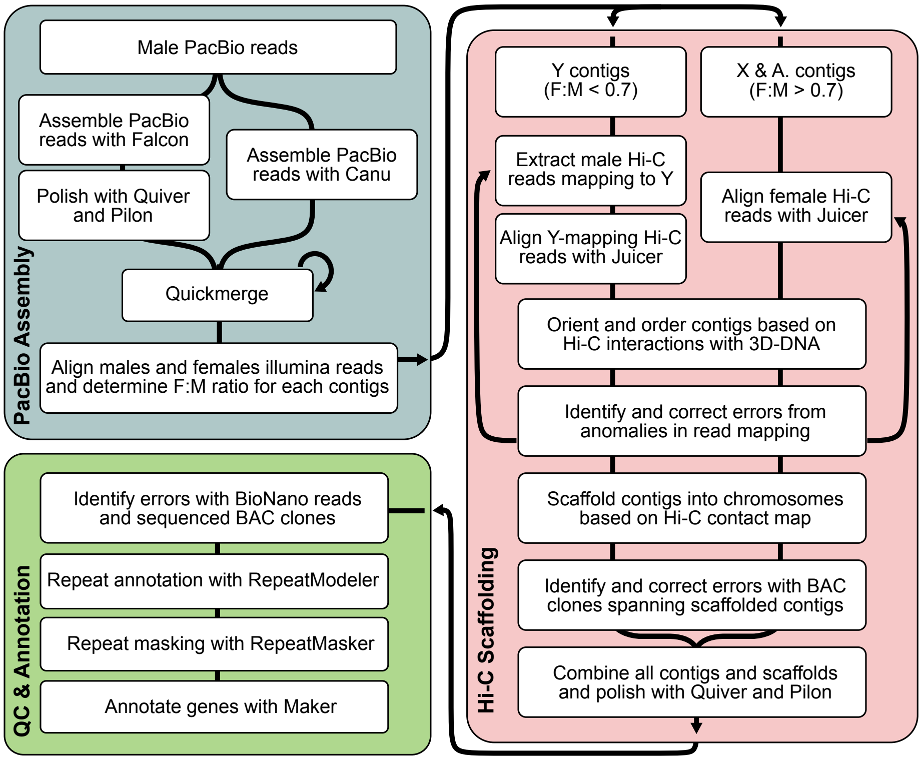

PacBio Assembly
Male PacBio reads
Assemble PacBio reads with Falcon
Assemble PacBio reads with Canu
Polish with Quiver and Pilon
Quickmerge
Align males and females illumina reads and determine F:M ratio for each contigs
Hi-C Scaffolding
Y contigs (F:M < 0.7)
X & A. contigs (F:M > 0.7)
Extract male Hi-C reads mapping to Y
Align Y-mapping Hi-C reads with Juicer
Align female Hi-C reads with Juicer
Orient and order contigs based on Hi-C interactions with 3D-DNA
Identify and correct errors from anomalies in read mapping
Scaffold contigs into chromosomes based on Hi-C contact map
Identify and correct errors with BAC clones spanning scaffolded contigs
Combine all contigs and scaffolds and polish with Quiver and Pilon
QC & Annotation
Identify errors with BioNano reads and sequenced BAC clones
Repeat annotation with RepeatModeler
Repeat masking with RepeatMasker
Annotate genes with Maker

Figure 2. Overview of the Genome assembly and annotation pipeline.

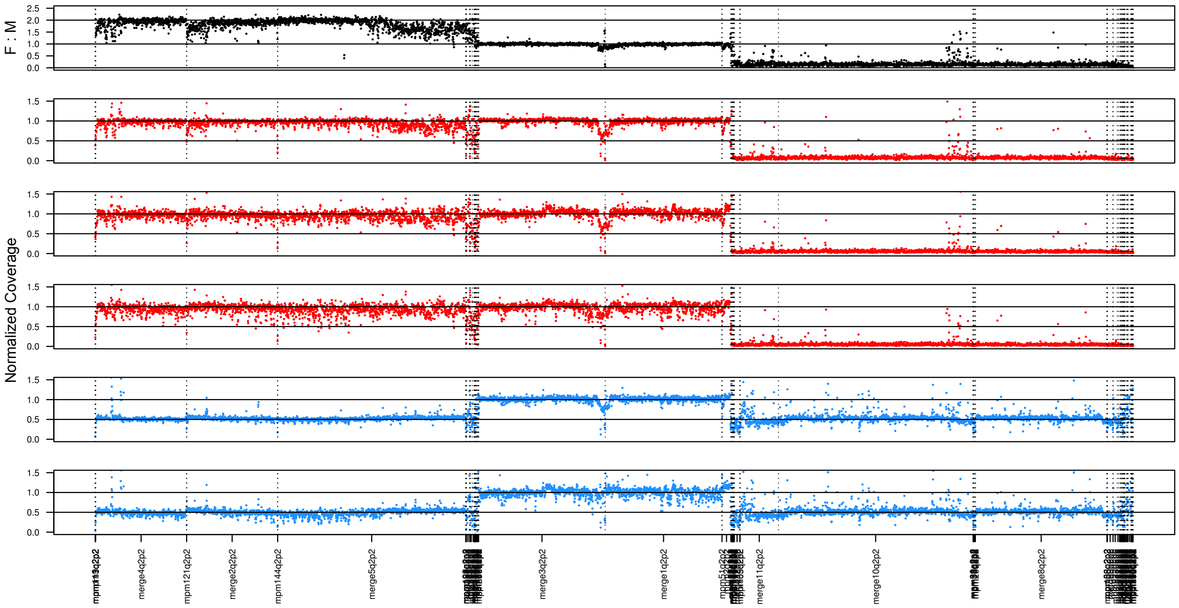


Figure 3. Normalized coverage across the genome for males (in blue), females (in red) and Female/Male (in black).

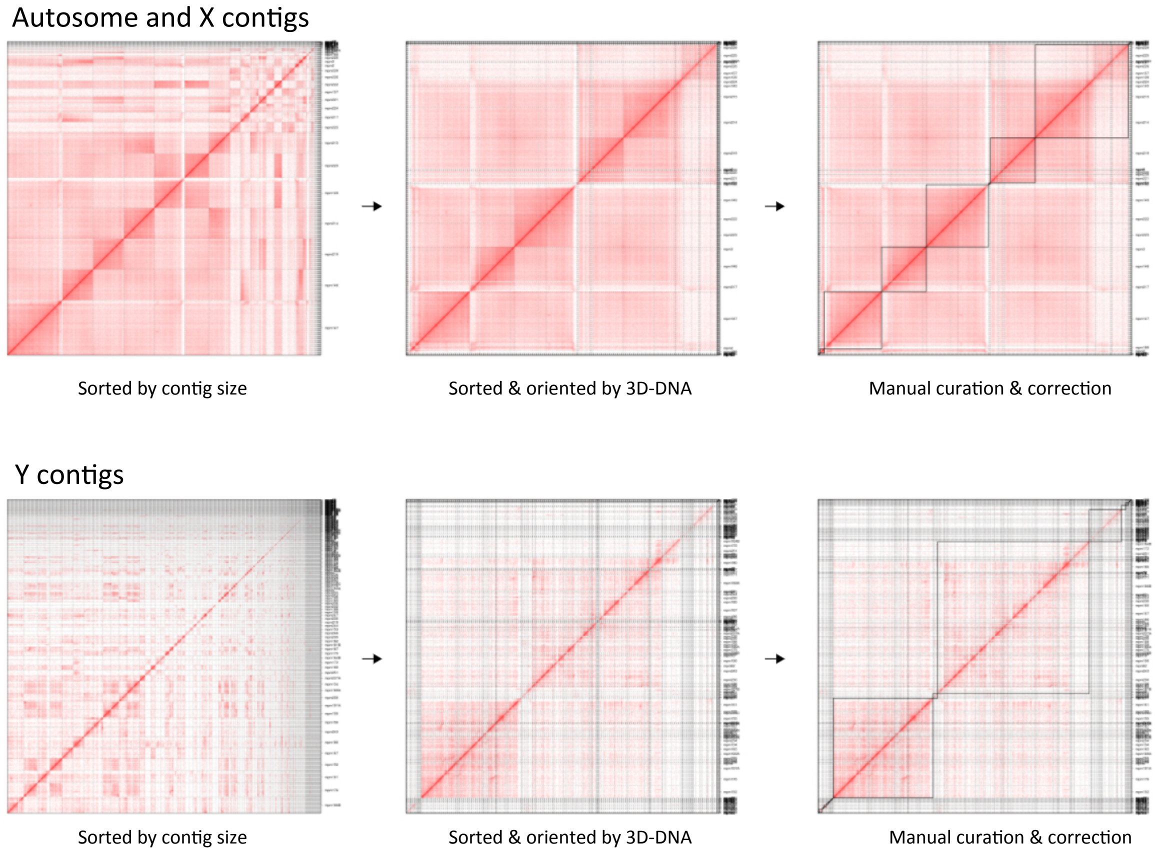


Figure 4. Scaffolding the genome using HiC interactions.

Autosomes and X-linked contigs were scaffolded separately from the Y/Neo-Y-linked contigs.

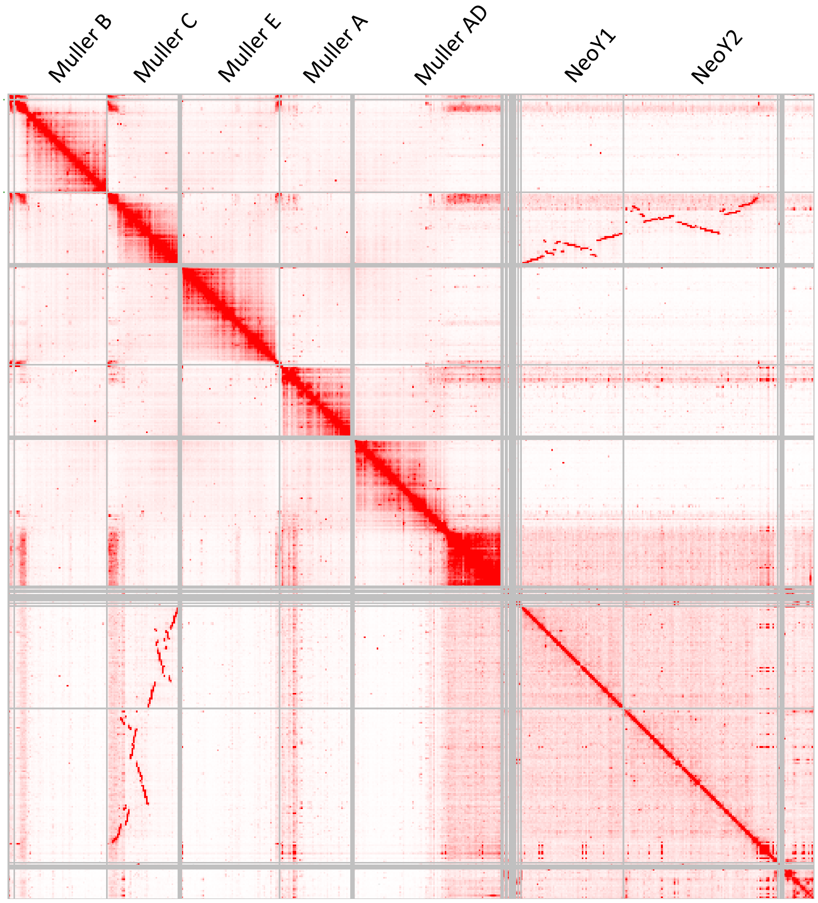


Figure 5. HiC interaction map of the de novo assembled *D.miranda* genome.

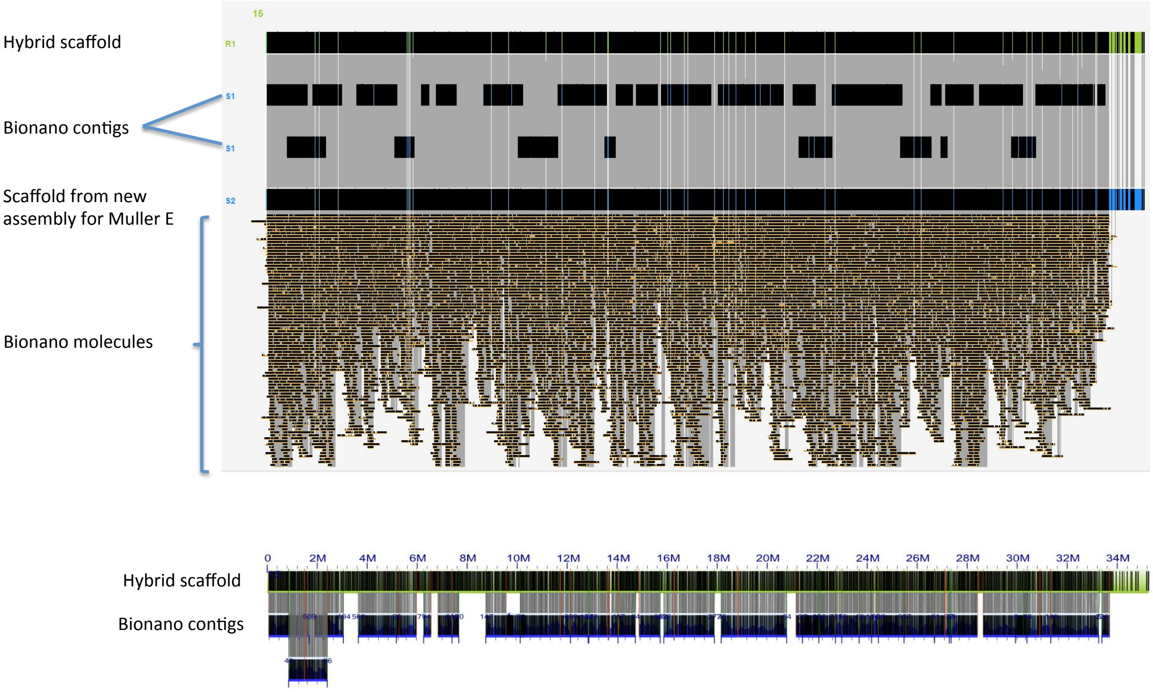


Figure 6. Alignment of Bionano contigs and Muller E chromosome from the new *D.miranda* genome assembly to the bionano hybrid scaffold created by merging the bionano optical map with the genome assembly.

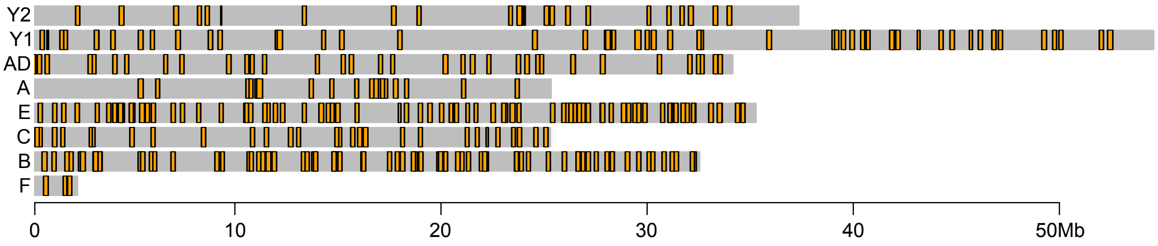


Figure 7. Alignment of BAC clones to the *denovo* genome assembly.

Grey bars represent chromosomes and orange boxes show the locations of the alignment of BAC clones.

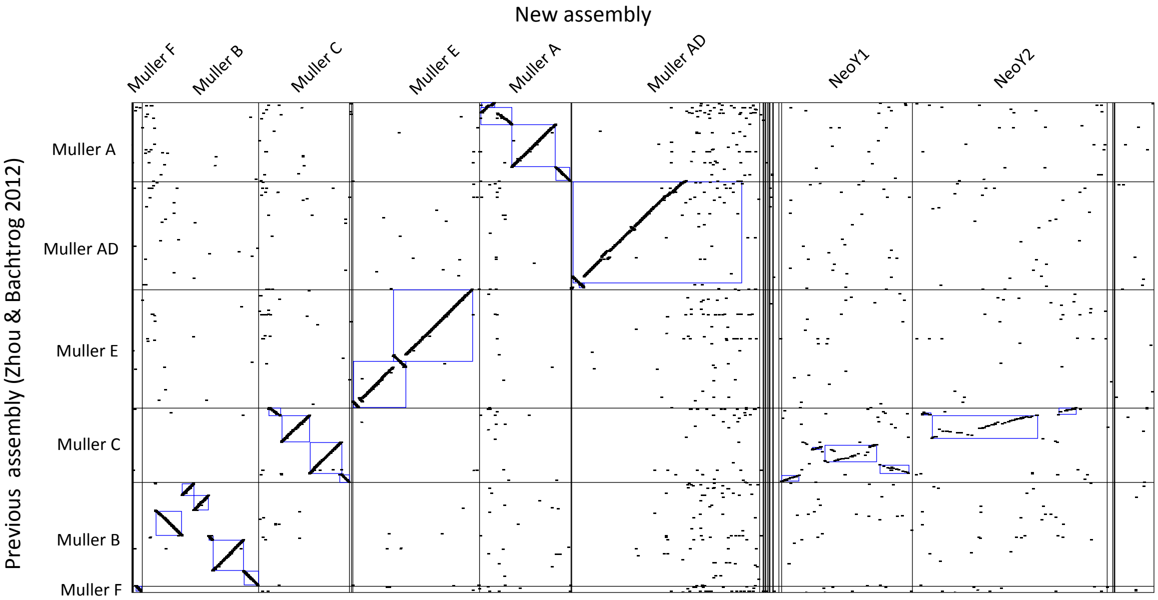


Figure 8. Dotplot showing the alignment between the old, previously published D.miranda genome assembly and the new *D.miranda* assembly.

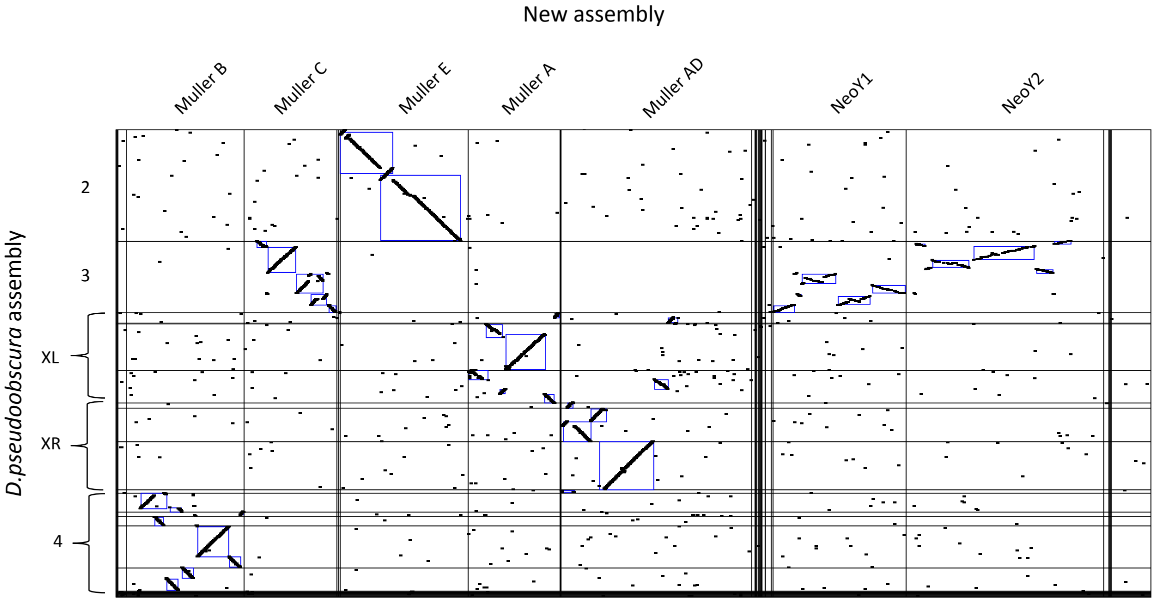


Figure 9. Dotplot showing the alignment between the previously published *D.pseudoobscura* genome assembly and the new *D.miranda* assembly.

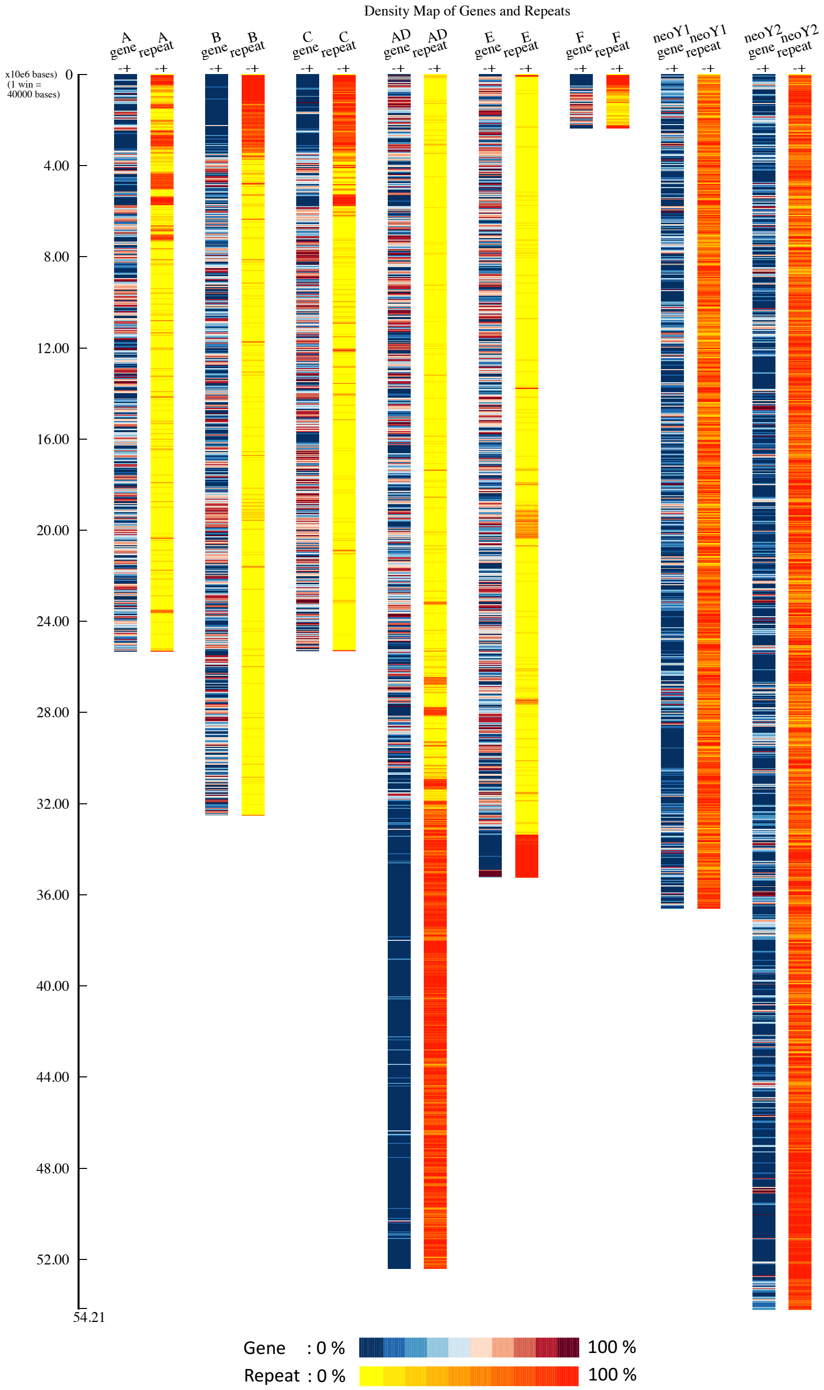


Figure 11. Gene and Repeat density across different chromosomes.

Regions in dark red indicate regions with high gene density and gene-poor regions are indicared in blue. Repeat rich regions are shown in red and regions with low repeat density are shown in yellow.

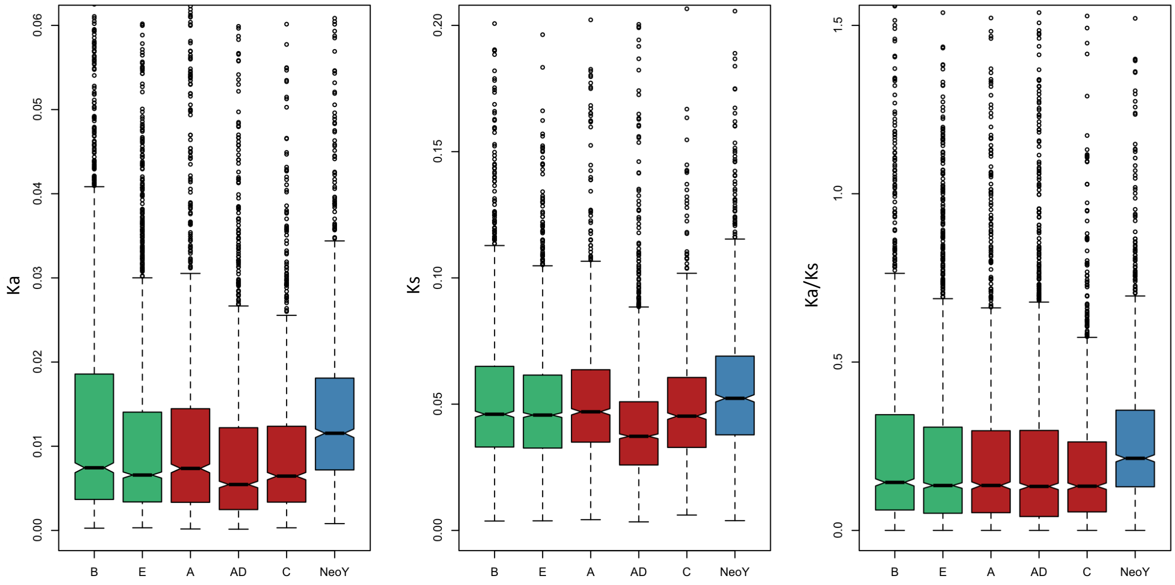
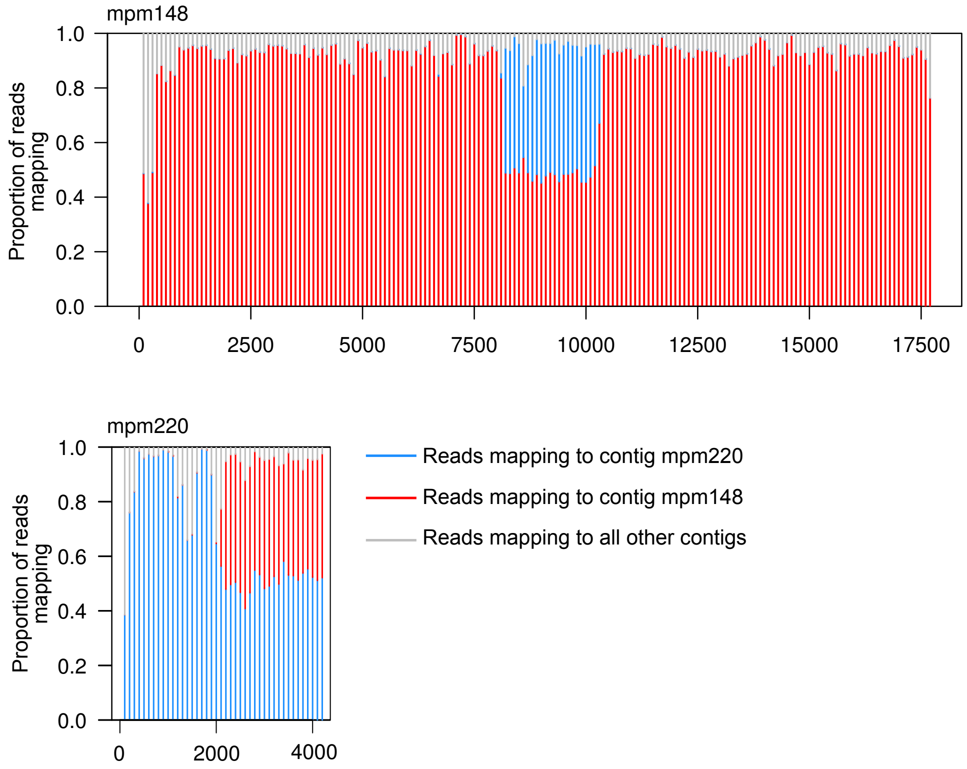


Figure 11. Rates of protein evolution between *D.pseudoobscura* and *D.miranda.*

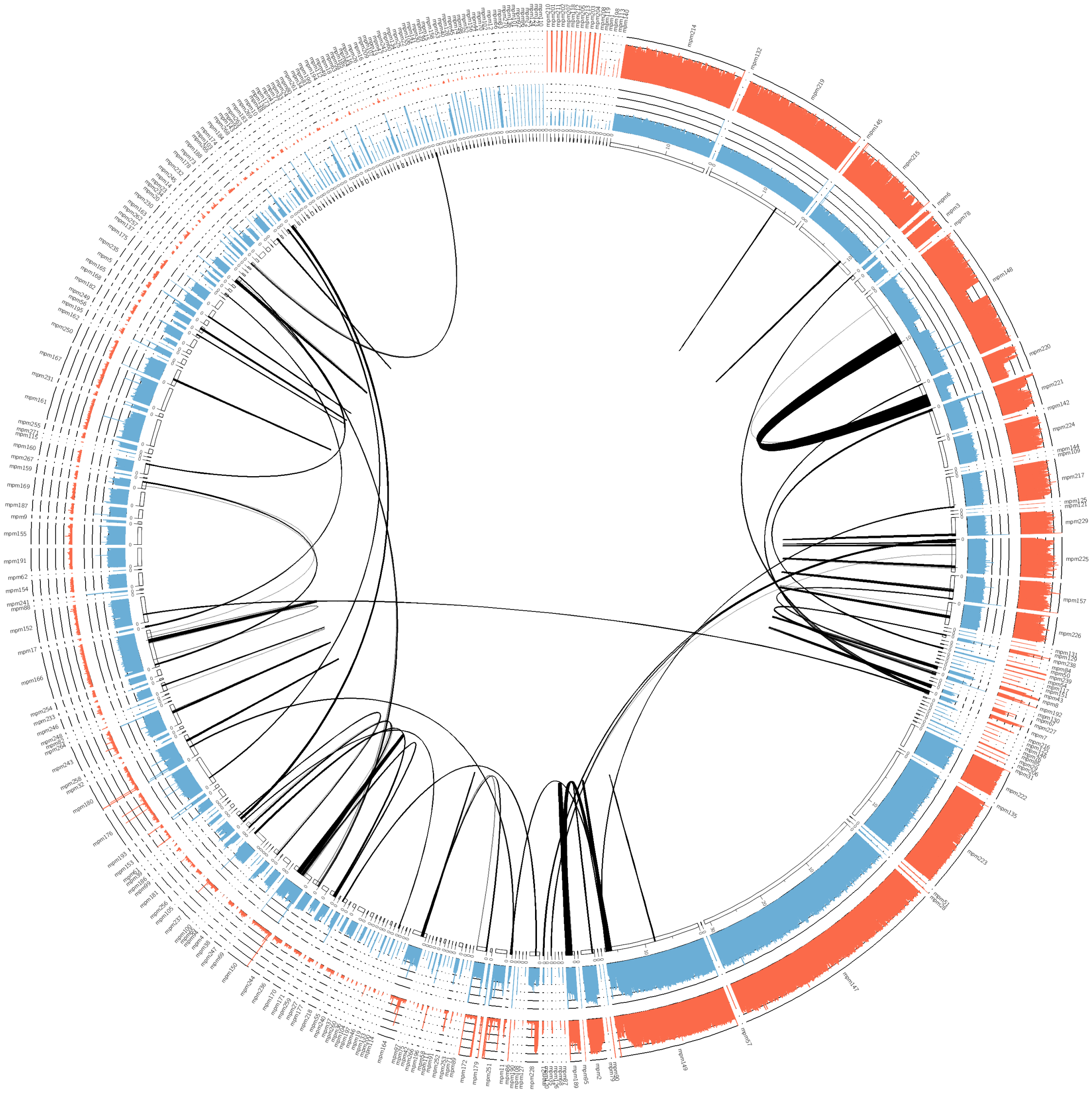
Different Muller Elements or chromosomes are indicated along the X axis.

Supplementary Figures



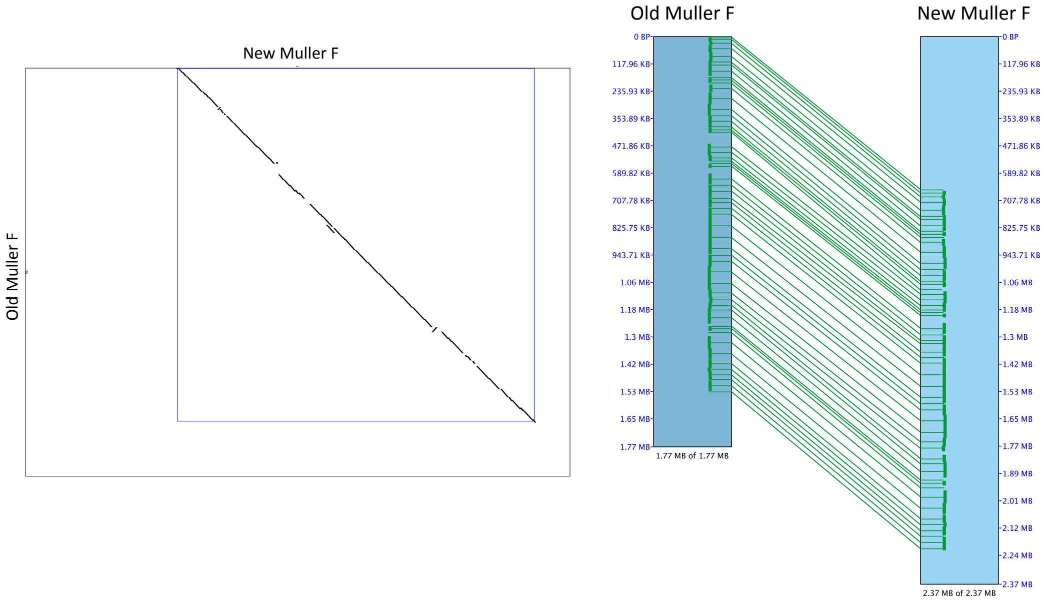
Supplementary Figure S1. Misassembly identification using illumina short reads alignments.

We aligned illumina reads to the assembly twice and identified >100kb regions where approximately half of the reads align to another region in the two mappings to identify potential misassemblies. One such example of misassembly is shown in this figure.

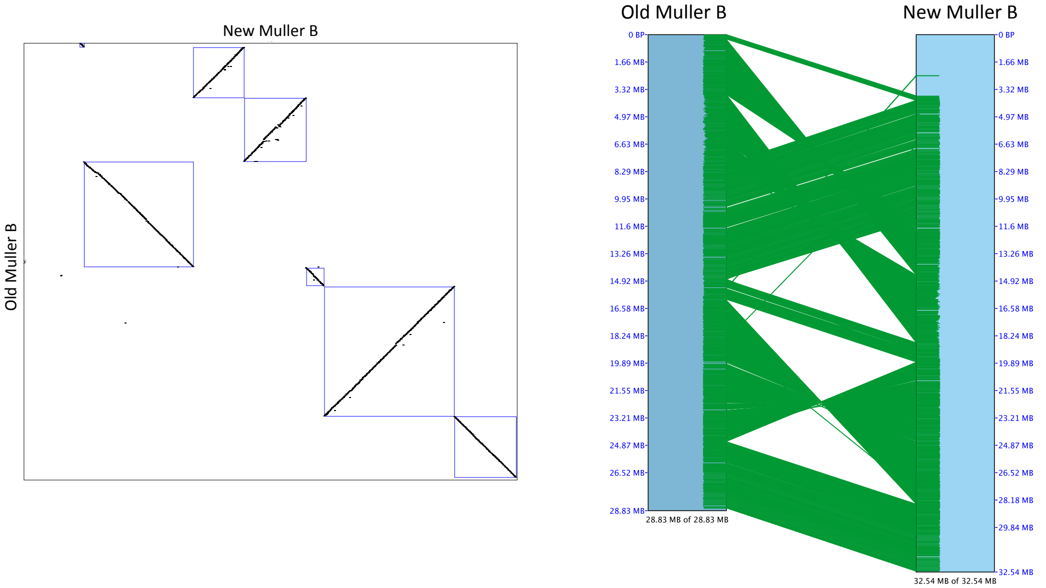


Supplementary Figure S2. Identifying falsely duplicated regions in the genome.

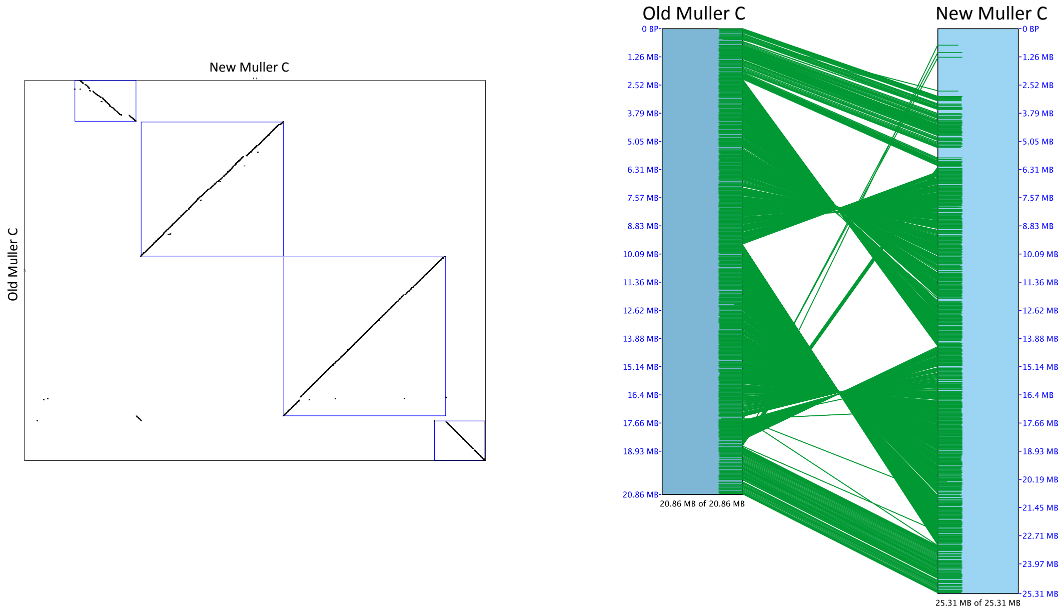
Circos plot of scaffolds in the genome assembly. Male genomic coverage is shown in blue and female coverage is shown in red. Black lines indicate regions that are duplicated in the genome. A duplication accompanied by a drop in coverage (to approximately half) indicates a misassembly due to erroneous duplication.



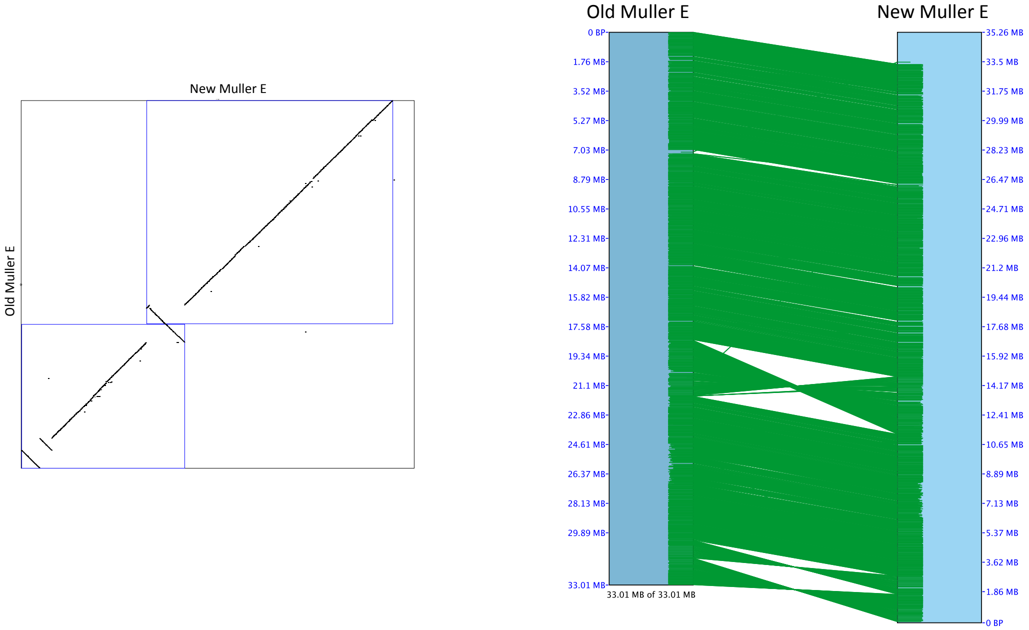
Supplementary Figure 3a. Alignment between the Muller F in the previous genome assembly and the new *D.miranda* genome assembly.



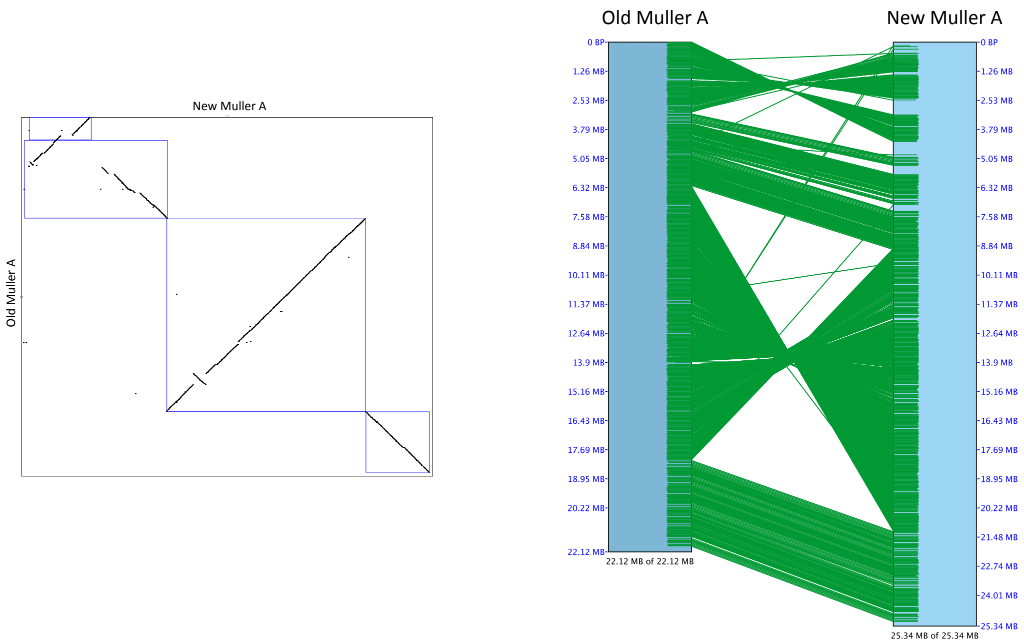
Supplementary Figure 3b. Alignment between the Muller B in the previous genome assembly and the new *D.miranda* genome assembly.



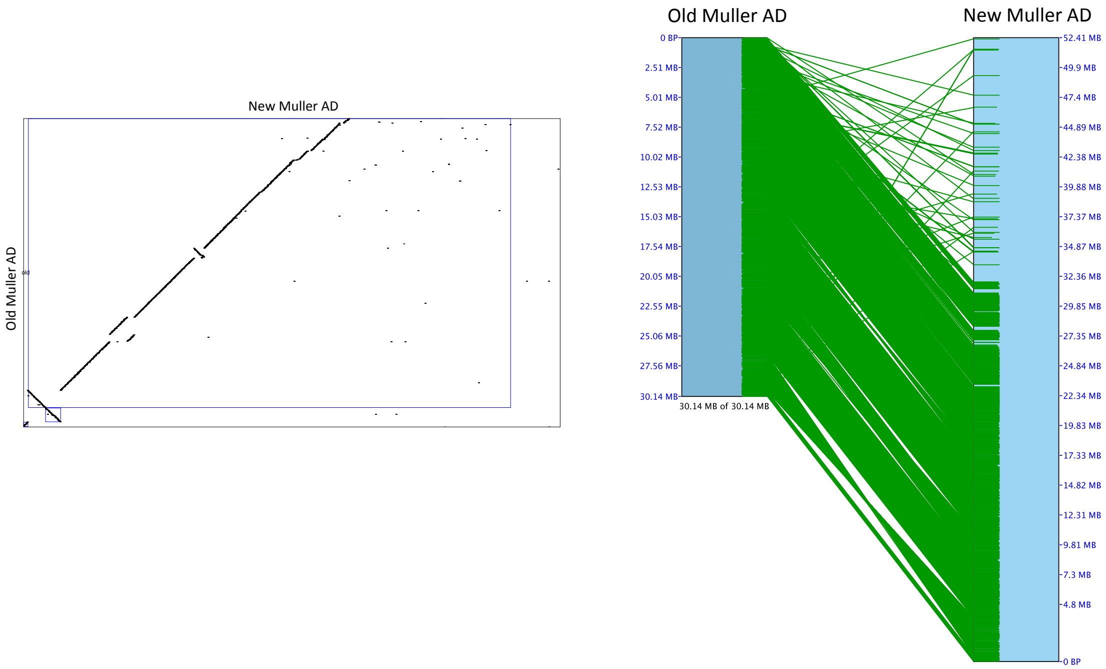
Supplementary Figure 3c. Alignment between the Muller C in the previous genome assembly and the new *D.miranda* genome assembly.



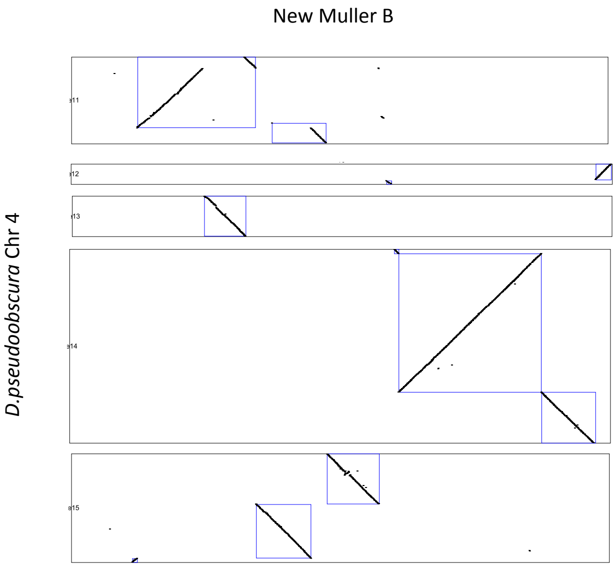
Supplementary Figure 3d. Alignment between the Muller E in the previous genome assembly and the new *D.miranda* genome assembly.



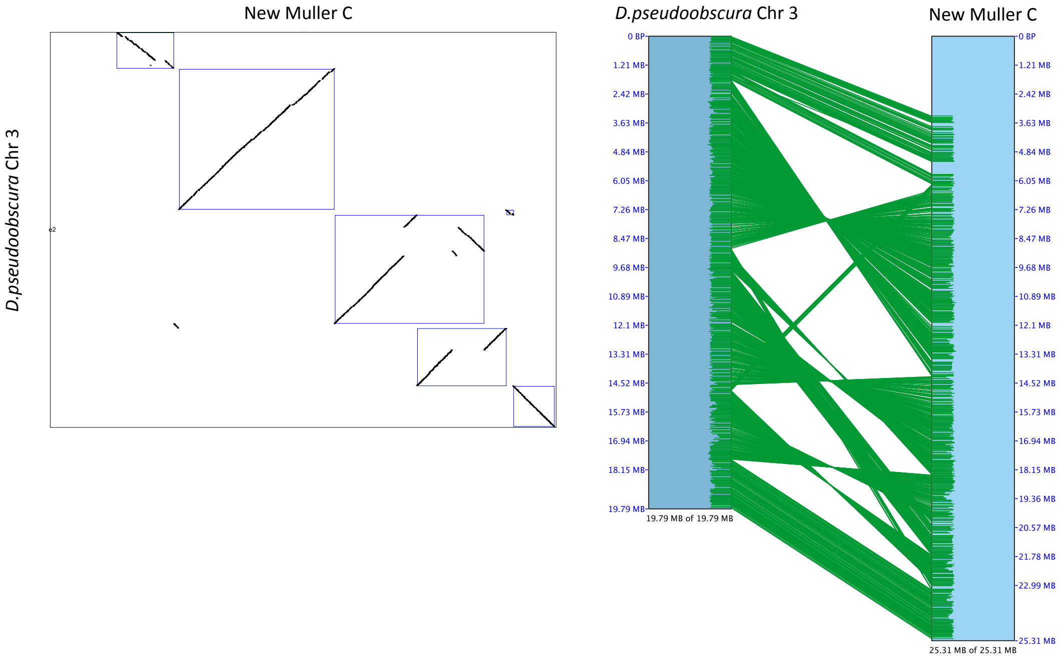
Supplementary Figure 3e. Alignment between the Muller A in the previous genome assembly and the new *D.miranda* genome assembly.



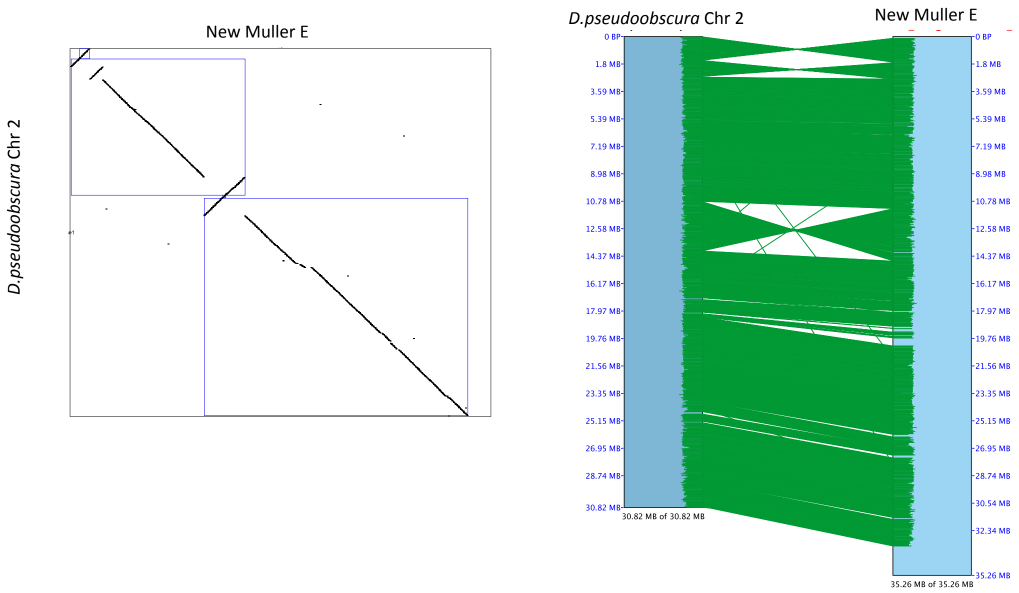
Supplementary Figure 3f. Alignment between the Muller AD in the previous genome assembly and the new *D.miranda* genome assembly.



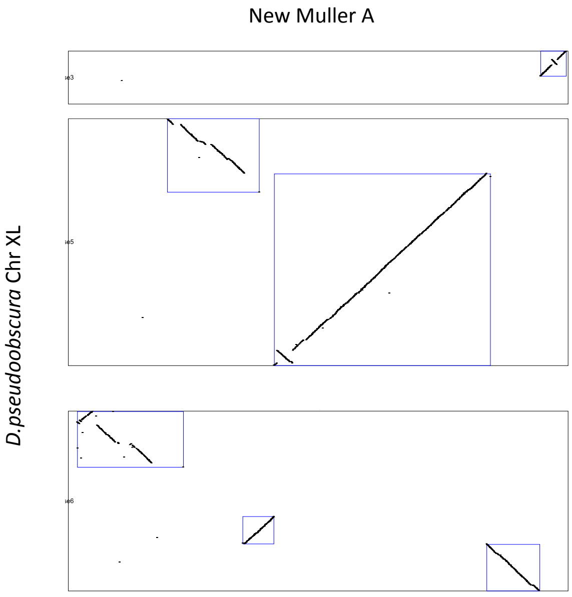
Supplementary Figure 4a. Alignment between Chr 4 (Muller B) in the *D.pseudoobscura* published assembly and the Muller B in the new *D.miranda* genome assembly.



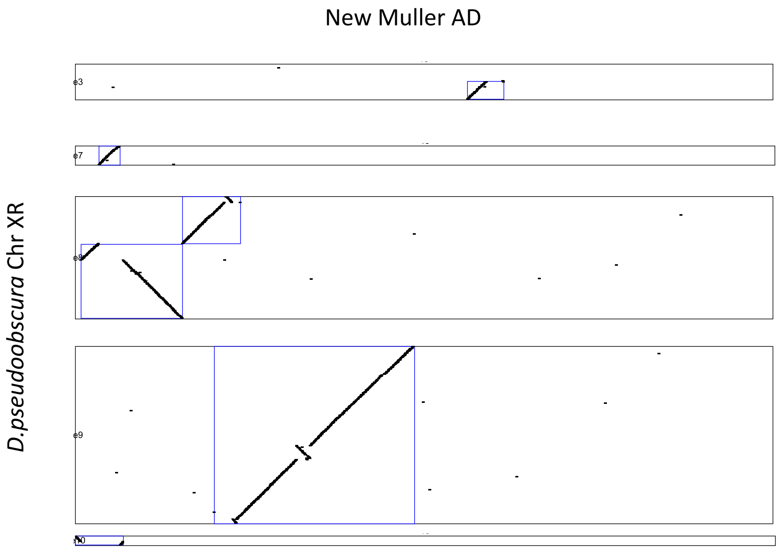
Supplementary Figure 4b. Alignment between Chr 3 (Muller C) in the *D.pseudoobscura* published assembly and the Muller C in the new *D.miranda* genome assembly.



Supplementary Figure 4c. Alignment between Chr 2 (Muller E) in the *D.pseudoobscura* published assembly and the Muller E in the new *D.miranda* genome assembly.



Supplementary Figure 4d. Alignment between Chr XL (Muller A) in the *D.pseudoobscura* published assembly and the Muller A in the new *D.miranda* genome assembly.



Supplementary Figure 4e. Alignment between Chr XR (Muller AD) in the *D.pseudoobscura* published assembly and the Muller AD in the new *D.miranda* genome assembly.

Supplementary tables

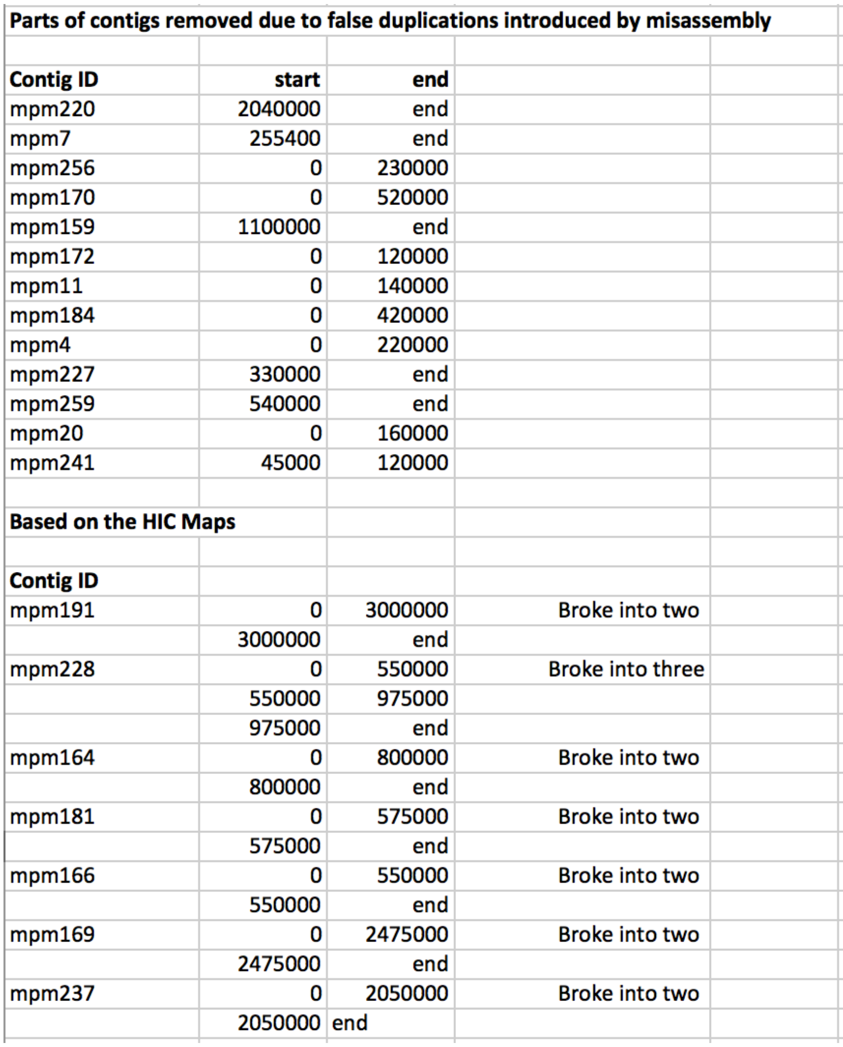
Supplementary Table S1. Overview of the datasets used.

Sequencing	sample	Sex	Number of reads	Comment
Illumina HiSeq	MSH22 female	Female	32801952 pairs	285 bp insert
Illumina HiSeq	MSH22 female	Female	8244007 pairs	640 bp insert
Illumina HiSeq	MSH22 female	Female	11461680 pairs	918 bp insert
Illumina HiSeq	MSH22 male	Male	12494994 pairs	626 bp insert
Illumina HiSeq	MSH22 male	Male	9724346 pairs	915 bp insert
Pacbio	MSH22 male	Male	2407465 reads	Read N50 17.1Kbp
BAC clones	MSH322 male	Male	365503123 pairs	Nextera fragmented (65-465 bp; peak 161 bp insert)
Hi-C	MSH22 male	Male	27990598 pairs	HpyCH4IV (4 cutter)
Hi-C	MSH22 female	Female	38433349 pairs	HpyCH4IV (4 cutter)

Bionano map

N50	Number of contigs	Cumulative length
0,.51Mb	401	173.84Mb

Supplementary Table S2. Manual curation and correction of the assembly based on illumina mapping and HiC interactions.



Parts of contigs removed due to false duplications introduced by misassembly			
Contig ID	**start**	**end**	
mpm220	2040000	end	
mpm7	255400	end	
mpm256	0	230000	
mpm170	0	520000	
mpm159	1100000	end	
mpm172	0	120000	
mpm11	0	140000	
mpm184	0	420000	
mpm4	0	220000	
mpm227	330000	end	
mpm259	540000	end	
mpm20	0	160000	
mpm241	45000	120000	
Based on the HIC Maps			
Contig ID			
mpm191	0	3000000	Broke into two
	3000000	end	
mpm228	0	550000	Broke into three
	550000	975000	
	975000	end	
mpm164	0	800000	Broke into two
	800000	end	
mpm181	0	575000	Broke into two
	575000	end	
mpm166	0	550000	Broke into two
	550000	end	
mpm169	0	2475000	Broke into two
	2475000	end	
mpm237	0	2050000	Broke into two
	2050000	end	

References

Bachtrog D., 2013 Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. **14**: 113–124.

Bachtrog D., Charlesworth B., 2002 Reduced adaptation of a non-recombining neo-Y chromosome. Nature **416**: 323–326.

Bachtrog D., Hom E., Wong K. M., Maside X., de Jong P., 2008 Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biol. **9**: R30.

Bachtrog D., Mank J. E., Peichel C. L., Kirkpatrick M., Otto S. P., Ashman T.-L., Hahn M. W., Kitano J., Mayrose I., Ming R., Perrin N., Ross L., Valenzuela N., Vamosi J. C., Tree of Sex Consortium, 2014 Sex determination: why so many ways of doing it? PLoS Biol. **12**: e1001899.

Bartolomé C., Charlesworth B., 2006 Rates and patterns of chromosomal evolution in Drosophila pseudoobscura and D. miranda. Genetics **173**: 779–791.

Bickhart D. M., Rosen B. D., Koren S., Sayre B. L., Hastie A. R., Chan S., Lee J., Lam E. T., Liachko I., Sullivan S. T., Burton J. N., Huson H. J., Nystrom J. C., Kelley C. M., Hutchison J. L., Zhou Y., Sun J., Crisà A., Ponce de León F. A., Schwartz J. C., Hammond J. A., Waldbieser G. C., Schroeder S. G., Liu G. E., Dunham M. J., Shendure J., Sonstegard T. S., Phillippy A. M., Van Tassell C. P., Smith T. P. L., 2017 Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat. Genet. **431**: 931.

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

Chakraborty M., Baldwin-Brown J. G., Long A. D., Emerson J. J., 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. **44**: e147.

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

Chin C.-S., Peluso P., Sedlazeck F. J., Nattestad M., Concepcion G. T., Clum A., Dunn C., O'Malley R., Figueroa-Balderas R., Morales-Cruz A., Cramer G. R., Delledonne M., Luo C., Ecker J. R., Cantu D., Rank D. R., Schatz M. C., 2016 Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods **13**: 1050–1054.

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

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

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

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

Ellison C. E., Bachtrog D., 2013 Dosage compensation via transposable element mediated rewiring of a regulatory network. Science **342**: 846–850.

Gramates L. S., Marygold S. J., Santos G. D., Urbano J.-M., Antonazzo G., Matthews B. B., Rey A. J., Tabone C. J., Crosby M. A., Emmert D. B., Falls K., Goodman J. L., Hu Y., Ponting L., Schroeder A. J., Strelets V. B., Thurmond J., Zhou P., the FlyBase Consortium, 2017 FlyBase at 25: looking to the future. Nucleic Acids Res. **45**: D663–D671.

Hoskins R. A., Smith C. D., Carlson J. W., Carvalho A. B., Halpern A., Kaminker J. S., Kennedy C., Mungall C. J., Sullivan B. A., Sutton G. G., Yasuhara J. C., Wakimoto B. T., Myers E. W., Celniker S. E., Rubin G. M., Karpen G. H., 2002 Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. **3**: RESEARCH0085.

Khost D. E., Eickbush D. G., Larracuente A. M., 2017 Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. Genome Res. **27**: 709–721.

Kim D., Langmead B., Salzberg S. L., 2015 HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**: 357–360.

Koren S., Walenz B. P., Berlin K., Miller J. R., Bergman N. H., Phillippy A. M., 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**: 722–736.

Korf I., 2004 Gene finding in novel genomes. BMC Bioinformatics **5**: 59.

Langmead B., Salzberg S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

Layer R. M., Chiang C., Quinlan A. R., Hall I. M., 2014 LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. **15**: R84.

Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754–1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

Löytynoja A., 2014 Phylogeny-aware alignment with PRANK. Methods Mol. Biol. **1079**: 155–170.

Mahajan S., Bachtrog D., 2017 Convergent evolution of Y chromosome gene content in flies. Nat Commun **8**: 785.

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

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

Schatz M. C., Delcher A. L., Salzberg S. L., 2010 Assembly of large genomes using second-generation sequencing. Genome Res. **20**: 1165–1173.

Skaletsky H., Kuroda-Kawaguchi T., Minx P. J., Cordum H. S., Hillier L., Brown L. G., Repping S., Pyntikova T., Ali J., Bieri T., Chinwalla A., Delehaunty A., Delehaunty K., Du H., Fewell G., Fulton L., Fulton R., Graves T., Hou S.-F., Latrielle P., Leonard S., Mardis E., Maupin R., McPherson J., Miner T., Nash W., Nguyen C., Ozersky P., Pepin K., Rock S., Rohlfing T., Scott K., Schultz B., Strong C., Tin-Wollam A., Yang S.-P., Waterston R. H., Wilson R. K., Rozen S., Page D. C., 2003 The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**: 825–837.

Smith A., Hubley R., *RepeatModeler Open-1.0*. httpwww.repeatmasker.org.

Smith A., Hubley R., Green P., *RepeatMasker Open-4.0*. httpwww.repeatmasker.org.

Soh Y. Q. S., Alföldi J., Pyntikova T., Brown L. G., Graves T., Minx P. J., Fulton R. S., Kremitzki C., Koutseva N., Mueller J. L., Rozen S., Hughes J. F., Owens E., Womack J. E., Murphy W. J., Cao Q., de Jong P., Warren W. C., Wilson R. K., Skaletsky H., Page D. C., 2014 Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell **159**: 800–813.

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

Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L., 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc **7**: 562–578.

Vicoso B., Charlesworth B., 2006 Evolution on the X chromosome: unusual patterns and processes. Nat. Rev. Genet. **7**: 645–653.

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

Zhang Z., Li J., Zhao X.-Q., Wang J., Wong G. K.-S., Yu J., 2006 KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics **4**: 259–263.

Zhou Q., Bachtrog D., 2012 Sex-specific adaptation drives early sex chromosome evolution in Drosophila. Science **337**: 341–345.

Zhou Q., Ellison C. E., Kaiser V. B., Alekseyenko A. A., Gorchakov A. A., Bachtrog D., 2013 The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. (PB Becker, Ed.). PLoS Biol. **11**: e1001711.

Chapter 4

Establishment of heterochromatin during early development in Drosophila

Shivani Mahajan & Doris Bachtrog

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

Abstract

Significant portions of eukaryotic genomes, including the Y chromosome, are heterochromatic, made up largely of repetitive sequences and possessing a distinctive chromatin structure associated with gene silencing. Heterochromatic regions have a high repeat content and are characterized by specific histone modifications, but the primary sequence elements that define specific chromosomal domains as preferred sites of heterochromatin assembly are not well understood. Here, we characterize the establishment of heterochromatin during early development in *Drosophila miranda*, a species that harbors a recently formed, partly heterochromatic neo-Y chromosome. We find that heterochromatin levels increase during the onset of zygotic genome activation, and males show a temporal delay in their establishment of heterochromatin relative to females. We find that transposable elements (TEs) that are inserted in euchromatic regions show spreading of heterochromatin, with the spreading signal being more pronounced for TEs that are being targeted by a higher number of maternally inherited piRNAs. This suggests that piRNAs are involved in initiating heterochromatin formation at repeats in Drosophila.

Introduction

In most animals, embryonic development is initially controlled solely by maternal proteins and transcripts (Newport and Kirschner 1982a; b), before zygotic transcription initiates (the maternal-to-zygotic transition). In *Drosophila melanogaster*, zygotic transcription begins about an hour into development (at the preblastoderm stage 2) and gradually increases; by the end of stage 4 (syncytial blastoderm), widespread zygotic transcription is observed (Pritchard and Schubiger 1996; Lécuyer *et al.* 2007). Zygotic genome activation is associated with massive remodeling of the chromatin architecture (Li *et al.* 2014). In particular, embryonic chromatin is in a relatively simple state at the end of stage 2, with undetectable levels of histone methylation marks, and low levels of histone acetylation at a relatively small number of loci. Histone acetylation increases in the syncytial blastoderm (stage 4), but it is not until stage 5 (cellularization of blastoderm) that nucleosome free regions and domains of histone methylation become widespread (**Figure 1**).

Concordant with genome-wide activation of zygotic expression, the embryo also has to assure that genomic regions are silenced whose transcription would be harmful (Haig 2016). In particular, large fractions of eukaryotic genomes consist of repetitive DNA, and activation of repeats could result in mobilization of repeats, causing insertional mutations and genomic instability (Hedges and Deininger 2007). Silencing of repeats is achieved through establishment of constitutive heterochromatin in all cells during early development at repetitive DNA at centromeres, telomeres, and along the Y chromosome (Peng and Karpen 2008; Swenson *et al.* 2016). An important questions remaining is how the decision is made to package a given region of the genome as heterochromatin.

Heterochromatin formation, and the boundary between heterochromatic and euchromatic domains is established during early development. In *D. melanogaster*, constitutive heterochromatin is not observed cytologically in the initial zygote, but emerges during blastoderm formation (stage 4; 2 hour embryo; (Vlassova *et al.* 1991; Lu *et al.* 1998)). Chromatin assembly during this period is prior to any significant zygotic transcription, and thus dependent on maternally loaded RNA and proteins. Analysis of an inducible reporter gene has found that silencing occurs at the onset of gastrulation (end of stage 6), about 1 hour after heterochromatin is visible cytologically (stage 6; (Lu *et al.* 1998). The extent of silencing increases as embryonic development progresses, and by stage 15 (dorsal closure, between 11.5 and 13 hours of development), silencing patterns reminiscent of those for third instar larvae are established (Lu *et al.* 1998). The piRNA pathway was shown to be required for the formation of heterochromatin in Drosophila when it is established at late blastoderm stage, but silent chromatin is then transmitted through cell divisions independent of the piRNA system (Gu and Elgin 2013), and chromatin patterns established in the late embryo appear to persist during differentiation (Rudolph *et al.* 2007). This suggests that the RNAi system plays a critical role in heterochromatin establishment, but may not be required to sustain chromatin states during development in Drosophila.

The study of heterochromatin has been slowed due to the difficulty of experimentally manipulating it. Its dearth of genes and abundance of sequence repeats and TEs mean that there are far fewer unique sequence tags that can aid sequencing and assembly efforts, and heterochromatin has largely been ignored from most model organism genome-sequencing efforts. Also, the highly evolved heterochromatic regions investigated in Drosophila and other organisms have been inherited as heterochromatin for millions of years, making it difficult to identify and distinguish the causative sites initiating heterochromatin formation from the highly repetitive background that is heterochromatic because of chromatin spreading.

Here, we investigate the establishment of heterochromatin during early embryogenesis in *D. miranda.* This species contains a neoY chromosome that has become heterochromatic only very recently in its evolutionary history. Heterochromatic islands on the neo-Y are surrounded by euchromatic, non-repetitive DNA, and these unique regions can be used as a backbone for sequence assembly of flanking heterochromatic regions. Also, the short time scale since when these neo-Y regions have been heterochromatic implies that few secondary changes have yet accumulated that dilute the initial DNA changes initiating heterochromatin formation.

Results

Quantification of H3K9me3 binding

To define the chromatin landscape before, during and after the establishment of heterochromatin, we collected *D. miranda* (MSH22) embryos for several stages – stage 3 (before the onset of heterochromatin formation, mid- and late stage 4 (when some heterochromatin is first detected), stage 5a (when heterochromatin becomes cytologically visible), stage 7 (when heterochromatin exerts its silencing effect)(**Figure 1A** and **Supplementary Table 1**). For this study, we used stage 4d embryos (2 males and 2 females) and stage 7 embryos (2 males and 2 females) only, for which we had sufficiently high sequencing coverage (**Supplementary Table 1**).

We adapted an ultra-low-input native ChIP protocol to assay histone modifications from a very small number of cells (10^3 to 10^6 cells per ChIP; (Brind'Amour *et al.* 2015)), therefore allowing us to perform ChIP-seq experiments in single embryos (**Figure 1B**). We employed a previously described normalization strategy (Bonhoure *et al.* 2014) to compare the genomic distribution and relative levels of chromatin marks across flies from different developmental stages. Specifically, we 'spiked in' a fixed amount of chromatin from stage 7 *D. melanogaster* embryos to each *D. miranda* chromatin sample prior to ChIP and sequencing. *D. melanogaster* chromatin served as an internal standard for the immunoprecipitation experiment, which along with the input, allowed us to obtain normalized H3K9me3 binding profiles and directly compare heterochromatin levels in different samples (see **Figure 1B** and Methods for details; see **Supplementary Figure 1** for the normalized ChIP signals from the *D. melanogaster* spike) .

Dramatic increase in global heterochromatin levels

Heterochromatin in Drosophila is mainly found at (peri)centromeric regions, telomeres, the dot chromosome, and the repetitive Y (Hoskins *et al.* 2002). Pericentromeric regions are repeat-rich and gene-poor and tend to show high levels of enrichment for silencing histone marks such as H3K9me3 and H3K9me2 (Chen *et al.* 2014)(Riddle *et al.* 2011). **Figure 2** shows genome-wide H3K9me3 binding profiles in male and female stage 4d and stage 7 embryos. Late stage 4d embryos already show a clear enrichment for H3K9me3 in the pericentromeric regions of the chromosomes, implying that H3K9me3 establishment starts at an earlier stage in development. By stage 7, the boundaries between euchromatin and heterochromatin are clearly defined. Globally there is a significant increase (Wilcoxon test p-value <2.16 10-16) in H3K9me3 binding in pericentromeric regions from stage 4d to stage 7 in both males and females (**Figure 3**).

D. miranda males have a large neo-Y chromosome (>100Mb) that retains considerable homology to the neo-X (which is about 30Mb in size) at certain loci but has also accumulated large amounts of repeats. Males thus contain more repetitive DNA that needs to be silenced, and repeat-rich regions in males may experience a delay in the establishment of heterochromatin compared to females. Genome wide subtraction plots between H3K9me3 binding in females versus males indicate that females indeed show a significantly higher enrichment of heterochromatin at the pericentromeres (Wilcoxon test p-value <2.16 10-16) than the males at stage 4d, at the onset of heterochromatin formation (**Figure 4**). By stage 7, males and females have attained similar levels of H3K9me3 binding at pericentromeres (**Figure 5**).

We observe a large heterochromatin island (>2Mb) on Muller E and one on Muller B. *In situ* experiments support the presence of such islands in *D. miranda* (unpublished data). These islands are mostly repetitive but also contain some genes and have large numbers of piRNA reads mapping to them and thus may contain piRNA clusters (**Supplementary Figure 2**). Such islands have not been observed in the *D. melanogaster* genome.

Interestingly, we were able to assemble and identify the pericentromeric region on the neo-Y chromosome, which shows higher enrichment for H3K9me3 compared to the rest of the chromosome (**Figure 2**).

Global heterochromatin changes at transposable elements

Stable heterochromatin is necessary to silence transposable elements (Hedges and Deininger 2007). We used the same normalization procedure that we used for the whole genome to calculate H3K9me3 enrichment signals across different repeats that were annotated in the *D. miranda* genome using RepeatModeler ((Smith and Hubley); see Chapter 3 for details on repeat annotation). At stage 4d more repeats are silenced in females compared to males (**Figure 6A**), while similar levels of H3K9me3 enrichment are observed in both sexes at stage 7. In both males (p-value Wilcoxon test < 2.16 10-16) and females (p-value Wilcoxon test < 1.377e-05), higher levels of H3K9me3 enrichment are observed at repeats at stage 7 compared to stage 4d (**Figure 6B**).

TE silencing at euchromatic insertion sites

The above analysis combines all genomic instances of a certain TE, irrespective of whether they are full length or fragmented, active or inactive, and irrespective of their insertion into hetero- or euchromatin. TEs in repeat-rich regions can be silenced either because they are directly targeted by the heterochromatin-inducing machinery, or because of heterochromatin spreading from adjacent genomic regions (Girton and Johansen 2008; Sienski *et al.* 2012). Heterochromatin at transposons inserted into euchromatic sites cannot be caused by spreading of heterochromatin along the chromosome and instead must have been initiated by directing modifying complexes *in situ*.

While reads derived from TEs may map to many different positions, individual repeat insertions in euchromatic regions can be analyzed by virtue of their unique flanking genomic surroundings. We annotated all repeat insertions within euchromatic areas of the genome (see Methods) that were larger than 1kb in size.

We first looked for evidence of spreading of the H3K9me3 mark at all 1743 euchromatic repeat insertions that were >1kb by investigating their flanking region (10Kb on either side of the insertion, divided into non-overlapping 100bp windows on each side). Metagene plots for all insertions showed clear evidence of 'spreading', wherein H3K9me3 enrichment is strongest in the immediate vicinity of the repeat insertion but attenuates further away from the TE (**Figure 7**). Similar to global patterns of H3K9me3 enrichment, we find the spreading signal to be stronger at the later stage of development (**Figure 7-9**). In addition, stage 4d females show a stronger signal for spreading than stage 4d males (**Figure 7-9**), consistent with the heterochromatin sink effect of the Y (Francisco and Lemos 2014).

If piRNAs help to target heterochromatin formation at euchromatic TE insertions, we expect TEs that are targeted by maternal piRNAs to show stronger levels of H3K9me3 enrichment (Brennecke *et al.* 2008). We identified repeat insertions in euchromatin that lie in regions of the genome that have large numbers of early embryo (0-1hr) piRNA reads mapping to them, by overlapping the top 5% genomic windows that have the highest number of piRNA counts with repeat insertions greater than 1kb. Metagene plots reveal that H3K9me3 enrichment is indeed stronger surrounding this subset of TEs that are targeted by maternally deposited piRNAs compared to all euchromatic TE insertions from the genome (**Figure 8** & **Figure 9**). This provides evidence that piRNAs play an important role in establishing stable heterochromatin in the genome (Brennecke *et al.* 2008).

Heterochromatin formation across neo-Y regions

Genes on the neo-Y chromosome are expressed at a lower level, both because of having malfunctional promoters and regulatory sequences, but also because of spreading of heterochromatin from adjacent regions (Zhou *et al.* 2013) (**Supplementary Figure 3**). To contrast global heterochromatin levels between the neo-Y and its homologous neo-X chromosome, we compared H3K9me3 enrichment for stage 7 embryos (**Figure 10**). As expected, we observe a higher global H3K9me3 enrichment on the neo-Y compared to the neo-X (Wilcoxon test p-value <2.16 10-16).

We investigated repeat insertions on the neo-Y to look for evidence of spreading of H3K9me3. We identified the top 1% 10kb genomic windows on the neo-Y chromosome that have the highest number of early embryo (0-1hr) piRNA reads mapping to them and identified 460 repeat insertions larger than 1kb that overlap these regions. H3K9me3 enrichment in 25kb regions flanking these insertions sites indeed shows the typical ‘spreading signal’, where the H3K9me3 enrichment attenuates away from the insertion site (**Figure 11**). As for global patterns of heterochromatin formation, we see higher levels of H3K9me3 enrichment in flanking regions of repeat insertions in stage 7 males compared to stage 4d males (**Figure 11**).

Discussion

Significant portions of eukaryotic genomes, including the Y chromosome, are heterochromatic, made up largely of repetitive sequences and possessing a distinctive chromatin structure associated with gene silencing. Heterochromatic regions have a high repeat content and are characterized by specific histone modifications, but the primary sequence elements that define specific chromosomal domains as preferred sites of heterochromatin assembly are not well understood. Recent studies suggest that small RNAs — possibly derived from transposable elements (TEs) — contribute to heterochromatin targeting. The recently formed neo-Y chromosomes of *D. miranda* are in the process of evolving altered chromatin structure: On the *D. miranda* neo-Y - which was formed about 1 MY ago - large segments have already acquired a heterochromatic appearance and TEs show a striking accumulation. About half of the neo-Y-loci have become non-functional, and most genes (~80%) are down-regulated from the neo-Y. This is supporting a link between heterochromatin formation and repetitive DNA, and its repressive effect on gene expression. *D. miranda* therefore provides a unique system to study the mechanisms and evolution of heterochromatin formation *in action* using evolutionarys approach.

Here we used a combination of comparative sequence and gene expression analysis, small RNA profiling and ChIP-seq experiments to map histone modifications associated with heterochromatin, to study the molecular basis of heterochromatin and how it evolves. We adapted a method to study H3K9me3 enrichment in single embryos, and found that heterochromatin starts to form very early during development, during the very rapid initial cell divisions. We found that heterochromatin formation is delayed in males, relative to females: both, global levels of heterochromatin enrichments, as well as the signature of heterochromatin formation at euchromatic TE insertions is less pronounced in male embryos relative to female embryos in the very early stages of development (i.e. stage 4d). Later in development (i.e. stage 7), these differences become less severe. Adult females have been shown to have more heterochromatin in autosomes and the X chromosome compared to adult males (Brown and Bachtrog 2014), and it has been proposed that the Y chromosome in males may act as a ‘sink’ for heterochromatic marks (Francisco and Lemos 2014). A similar phenomenon has been observed for the establishment of dosage compensation in *D. melanogaster* vs. *D. pseudoobscura* vs. *D. miranda* (Lott *et al.* 2014). In these species, roughly 20% vs. 40% vs. 60% of the genome is X-linked, and the onset of dosage compensation occurs at a later stage in development in *D. miranda* vs. *D.*

pseudoobscura, and the earliest in *D. melanogaster* (Lott *et al.* 2014). Thus, a larger fraction of the genome being X-linked may lead to the delay observed across these species.

Consistent with small RNAs being involved in heterochromatin targeting, we find that euchromatic TE insertions that are targeted by large numbers of maternally inherited piRNAs show higher levels of heterochromatin spreading during early embryogenesis. We also show that the increased repeat content of the neo-Y is associated with higher levels of heterochromatin formation relative to the neo-X.

Interestingly, out of the 46 embryos collected for this study, 5 were found to be aneuploids (**Supplementary Figure 4**). This high rate of aneuploidy (>10%) has not been observed in *D. melanogaster* embryos that were used as spike-in. *D. miranda* males have an unusual karyotype, due to the recent fusion of an autosome to the Y chromosome, which led to the formation of the neo-Y chromosome. The two X chromosomes in males (XL/XR and the neo-X) and the Y/neo-Y chromosome form a trivalent in male meiosis, and this may lead to problems in correct segregation of chromosomes (Macknight and COOPER 1944; COOPER 1946). It will be of interest to investigate if this high rate of aneuploidy is due to frequent problems in male meiosis in *D. miranda*.

Materials and methods

Embryo collection

Flies from the *D. miranda* MSH22 strain kept at 18°C and *D. melanogaster* Oregon-R strain kept at 25°C were used for this study. Molasses plates were prepped with yeast paste and flies were allowed to lay on them for 15 min (for *D. melanogaster*) or 30 minutes (for *D. miranda*). Embryos were then aged, washed, dechorionated with bleach, and staged live under a light microscope for 10 mins (*D. melanogaster*) or 15 mins (*D. miranda*). Following visual confirmation of the stages, embryos were flash frozen in liquid nitrogen.

Chip-seq experiments

Embryos were homogenized using a pipette tip and ChIP-seq was performed on single embryos using the ULI-NChIP protocol described in (Brind'Amour *et al.* 2015) with a few modifications. First, chromatin was digested at 21°C using micrococcal nuclease (MNase) (New England Biolabs) for 7.5 minutes. DNA from *D. melanogaster* stage 7 embryos (1 embryo for upto 4 *D. miranda* embryos, 2 or more pooled embryos when more than 4 experiments were performed at the same time) was added to the DNA from *D. miranda* staged single embryos, such that the spike DNA made up approximately 20% of the total chromatin in each experiment. 10% of the chromatin was used as input in each experiment and the rest was incubated for 2-6 hours with Dynabeads Protein G (Invitrogen). The H3K9me3 antibody (Diagenode, 1.65 ug/ul) was first incubated with Dynabeads Protein G for >3 hours to bind the antibody to the beads and then added to the chromatin (0.25 ul per embryo) and incubated overnight. A low-salt buffer and a high-salt buffer were

then used to wash the chromatin-antibody-bead complexes, from which DNA was eluted by shaking at 65°C for 1-1.5 hours. A mixture of phenol, chloroform and isomayl alcohol was used to extract DNA from both the ChIP and input samples.
Agencourt AmpureXP beads were then used to clean the DNA and libraries were prepared using the ThruPLEX DNA-seq kit (Rubicon). Two more rounds of AmpureXP bead cleanups were performed on the libraries before sequencing. Samples were sequenced at the Vincent J. Coates Genomic Sequencing Laboratory at UC Berkeley (100bp paired-end sequencing).

In total, we sequenced 46 embryos from various embryonic stages, 22 females, 19 males and 5 aneuploids. For this study, we eliminated samples that had low sequencing coverage. However, some of these samples have good quality libraries and can be sequenced to higher coverage. For some of the libraries, most of the reads were PCR duplicates, adaptors or bacterial contamination. This is particularly true for the very early stage embryos from which extremely small amounts of DNA is obtained during the ChIP.

Determining the sex of the Embryos
Paired end reads from the input for each experiment, were trimmed and aligned using bowtie2 (Langmead and Salzberg 2012) to our unpublished de novo *D. miranda* genome assembly (in prep), which was divided into 10kb non-overlapping windows. BEDtools (Quinlan and Hall 2010) was then used to calculate the coverages for each chromosome. The median coverage of the autosomes was then used to normalize the coverage of each chromosome. Males have a single copy of the X chromosome, and hence the X chromosome in males has half the genomic coverage compared to the autosomes. Females have two copies of the X and have the same genomic coverage for the X chromosome and the autosomes.

Normalization procedure
We used a modified version of the normalization procedure described in (Bonhoure *et al.* 2014) to normalize the ChIP data. Briefly, for both the sample (*D. miranda*) and spike (*D. melanogaster*), we divided the genome into 10kb windows. Paired-end reads for the input and ChIP were trimmed and aligned to the two genomes using bowtie2 (Langmead and Salzberg 2012). For *D. miranda* embryos, reads were aligned to sex-specific genomes. We calculated the number of reads mapping to each genomic window using BEDtools (Quinlan and Hall 2010). We then normalized the samples (*D. miranda*) and spikes (*D. melanogaster*) separately for sequencing depth. To do this, for each experiment, we calculated the total number of read counts of all genomic windows for the ChIPs. We then calculated the median of total counts and scaled the counts of all genomic windows so that their total count equaled this median value. The inputs were scaled to the same total count as the ChIPs.

The spike chromatin serves as an internal standard for the Immunoprecipitation experiments. Since the same amount of spike chromatin was added to each sample, any differences in ChIP signals of the spike chromatin are due to technical variations in the experiments. The spike chromatin can, therefore, be used to calculate normalization factors to adjust for ChIP efficiency in different experiments.

The input for each experiment gives us an expectation of background/non-specific signal. To calculate the scaling factor, for each of the spikes (*D. melanogaster*) in different experiments, we calculated the mean of the positive residuals of the linear regression of ChIP on input. The residuals reflect the difference between the observed and predicted values, and hence only positive residuals were used since we are interested in regions that are enriched for the histone mark compared to the input. We then calculated the mean of the means of residuals of all experiments and computed the normalization factor for each experiment as the mean of residuals divided by the mean of means of residuals.

For each of the samples (*D. miranda*), we calculated the positive residuals of the regression of ChIP on input and divided them by the normalization factor. We then calculated the Log_2 ChIP enrichment signals according to the estimator described in (Bonhoure *et al.* 2014).

H3K9me3 enrichment at repeats

Paired-end reads for both ChIP and input were mapped to a *de novo D. miranda* repeat annotation built using RepeatModeler (Smith and Hubley). The number of reads mapping to each repeat was calculated using BEDtools (Quinlan and Hall 2010). Normalized ChIP signals were computed similarly as described previously using the input and the *D. melanogaster* spike chromatin.

Identifying repeat insertions in the genome

RepeatMasker (Smith *et al.*) was used to mask repeats in the genome using a *de novo* repeat library built using the software RepeatModeler (Smith and Hubley). The gff file produced by RepeatMasker was used to identify the locations of repeat insertions in the genome. A set of 1743 insertions, greater than 1kb in size, were investigated for 'spreading' of H3K9me3. To identify repeats that may be targeted by piRNAs, we first aligned piRNA reads from early embryos to the *D. miranda* genome divided into non-overlapping 10kb windows using bowtie2 (Langmead and Salzberg 2012) and the read count for each window was calculated using BEDtools (Quinlan and Hall 2010). Windows having large number of piRNA reads mapping to them were identified and repeat insertions overlapping the top 5% windows were classified as being targeted by piRNAs. Regions 25kb and 50kb upstream and downstream of these insertions were divided into 125bp or 100bp bins and BEDtools (Quinlan and Hall 2010) was used to compute the number of ChIP and input reads mapping to these bins followed by the normalization procedure described previously to quantify H3K9me3 enrichment.
Similarly, 460 insertions were identified on the neo-Y chromosome and H3K9me3 enrichment was quantified in regions that are 25Kb upstream and downstream of the insertion site using 125 bp bins (200 bins on each side of the insertions).

References

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

Brennecke J., Malone C. D., Aravin A. A., Sachidanandam R., Stark A., Hannon G. J., 2008 An epigenetic role for maternally inherited piRNAs in transposon silencing. Science **322**: 1387–1392.

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

Brown E. J., Bachtrog D., 2014 The chromatin landscape of Drosophila: comparisons between species, sexes, and chromosomes. Genome Res. **24**: 1125–1137.

Chen Z.-X., Sturgill D., Qu J., Jiang H., Park S., Boley N., Suzuki A. M., Fletcher A. R., Plachetzki D. C., FitzGerald P. C., Artieri C. G., Atallah J., Barmina O., Brown J. B., Blankenburg K. P., Clough E., Dasgupta A., Gubbala S., Han Y., Jayaseelan J. C., Kalra D., Kim Y.-A., Kovar C. L., Lee S. L., Li M., Malley J. D., Malone J. H., Mathew T., Mattiuzzo N. R., Munidasa M., Muzny D. M., Ongeri F., Perales L., Przytycka T. M., Pu L.-L., Robinson G., Thornton R. L., Saada N., Scherer S. E., Smith H. E., Vinson C., Warner C. B., Worley K. C., Wu Y.-Q., Zou X., Cherbas P., Kellis M., Eisen M. B., Piano F., Kionte K., Fitch D. H., Sternberg P. W., Cutter A. D., Duff M. O., Hoskins R. A., Graveley B. R., Gibbs R. A., Bickel P. J., Kopp A., Carninci P., Celniker S. E., Oliver B., Richards S., 2014 Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. **24**: 1209–1223.

COOPER K. W., 1946 The mechanism of non-random segregation of sex chromosomes in male Drosophila miranda. Genetics **31**: 181–194.

Francisco F. O., Lemos B., 2014 How Do Y-Chromosomes Modulate Genome-Wide Epigenetic States: Genome Folding, Chromatin Sinks, and Gene Expression. Journal of Genomics **2**: 94–103.

Girton J. R., Johansen K. M., 2008 Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. Adv. Genet. **61**: 1–43.

Gu T., Elgin S. C. R., 2013 Maternal depletion of Piwi, a component of the RNAi system, impacts heterochromatin formation in Drosophila. (J Brennecke, Ed.). PLoS Genet. **9**: e1003780.

Haig D., 2016 Transposable elements: Self-seekers of the germline, team-players of the soma. Bioessays **38**: 1158–1166.

Hedges D. J., Deininger P. L., 2007 Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res **616**: 46–59.

Hoskins R. A., Smith C. D., Carlson J. W., Carvalho A. B., Halpern A., Kaminker J. S., Kennedy C., Mungall C. J., Sullivan B. A., Sutton G. G., Yasuhara J. C., Wakimoto B. T., Myers E. W., Celniker S. E., Rubin G. M., Karpen G. H., 2002 Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol. **3**: RESEARCH0085.

Langmead B., Salzberg S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

Lécuyer E., Yoshida H., Parthasarathy N., Alm C., Babak T., Cerovina T., Hughes T. R., Tomancak P., Krause H. M., 2007 Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell **131**: 174–187.

Li X.-Y., Harrison M. M., Villalta J. E., Kaplan T., Eisen M. B., 2014 Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. Elife **3**: e1003428.

Lott S. E., Villalta J. E., Zhou Q., Bachtrog D., Eisen M. B., 2014 Sex-specific embryonic gene expression in species with newly evolved sex chromosomes. (HS Malik, Ed.). PLoS Genet. **10**: e1004159.

Lu B. Y., Ma J., Eissenberg J. C., 1998 Developmental regulation of heterochromatin-mediated gene silencing in Drosophila. Development **125**: 2223–2234.

Macknight R. H., COOPER K. W., 1944 The Synapsis of the Sex Chromosomes of Drosophila Miranda in Relation to Their Directed Segregation. Proc. Natl. Acad. Sci. U.S.A. **30**: 384–387.

Newport J., Kirschner M., 1982a A major developmental transition in early Xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage. Cell **30**: 675–686.

Newport J., Kirschner M., 1982b A major developmental transition in early Xenopus embryos: II. Control of the onset of transcription. Cell **30**: 687–696.

Peng J. C., Karpen G. H., 2008 Epigenetic regulation of heterochromatic DNA stability. Curr. Opin. Genet. Dev. **18**: 204–211.

Pritchard D. K., Schubiger G., 1996 Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio. Genes Dev. **10**: 1131–1142.

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

Riddle N. C., Minoda A., Kharchenko P. V., Alekseyenko A. A., Schwartz Y. B., Tolstorukov M. Y., Gorchakov A. A., Jaffe J. D., Kennedy C., Linder-Basso D., Peach S. E., Shanower G., Zheng H., Kuroda M. I., Pirrotta V., Park P. J., Elgin S. C. R., Karpen G. H., 2011 Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. **21**: 147–163.

Rudolph T., Yonezawa M., Lein S., Heidrich K., Kubicek S., Schäfer C., Phalke S., Walther M., Schmidt A., Jenuwein T., Reuter G., 2007 Heterochromatin formation in Drosophila is initiated through active removal of H3K4 methylation by the LSD1 homolog SU(VAR)3-3. Mol. Cell **26**: 103–115.

Sienski G., Dönertas D., Brennecke J., 2012 Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell **151**: 964–980.

Smith A., Hubley R., *RepeatModeler Open-1.0*. httpwww.repeatmasker.org.

Smith A., Hubley R., Green P., *RepeatMasker Open-4.0*. httpwww.repeatmasker.org.

Swenson J. M., Colmenares S. U., Strom A. R., Costes S. V., Karpen G. H., 2016 The composition and organization of Drosophila heterochromatin are heterogeneous and dynamic. Elife **5**: 1445.

Vlassova I. E., Graphodatsky A. S., Belyaeva E. S., Zhimulev I. F., 1991 Constitutive heterochromatin in early embryogenesis of Drosophila melanogaster. Mol. Gen. Genet. **229**: 316–318.

Zhou Q., Ellison C. E., Kaiser V. B., Alekseyenko A. A., Gorchakov A. A., Bachtrog D., 2013 The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. (PB Becker, Ed.). PLoS Biol. **11**: e1001711.

Figures

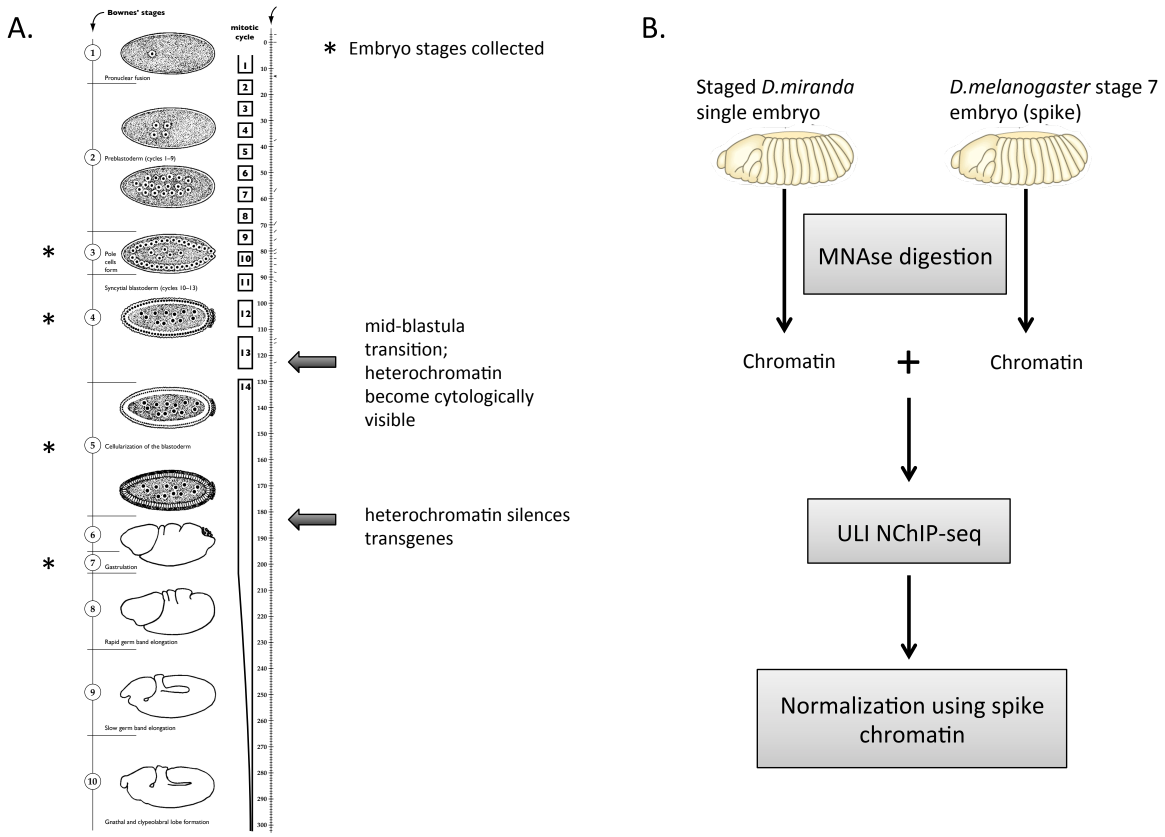

A.
Bownes' stages
mitotic cycle
* Embryo stages collected
1
Pronuclear fusion
2
Preblastoderm (cycles 1–9)
*
3
Pole cells form
Syncytial blastoderm (cycles 10–13)
*
4
*
5
Cellularization of the blastoderm
6
*
7
Gastrulation
8
Rapid germ band elongation
9
Slow germ band elongation
10
Gnathal and clypeolabral lobe formation
1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300
mid-blastula transition; heterochromatin become cytologically visible
heterochromatin silences transgenes
B.
Staged *D.miranda* single embryo
D.melanogaster stage 7 embryo (spike)
MNAse digestion
Chromatin
+
Chromatin
ULI NChIP-seq
Normalization using spike chromatin

Figure 1. A. Embryonic developmental stages for which samples were collected. B. Experimental design (Image 1A taken from http://kirschner.med.harvard.edu/files/bionumbers/Timetable%20of%20Drosophila%20Early%20Development.pdf).

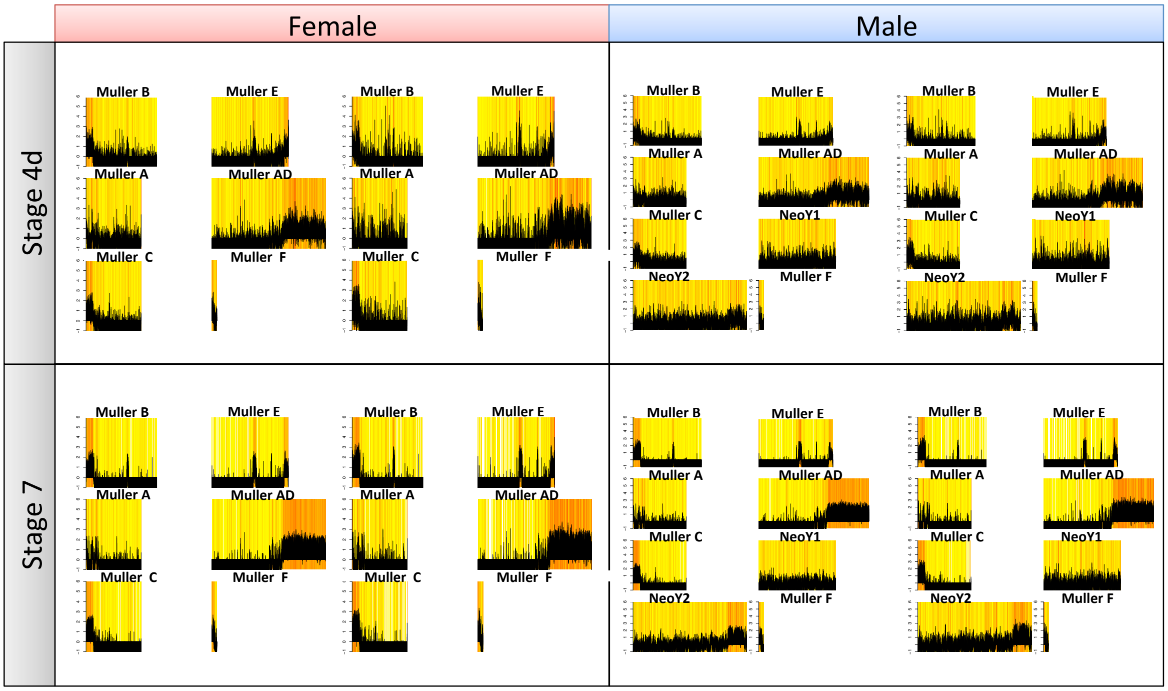


Figure 2. Genome wide H3K9me3 binding profiles in stage 4d and stage 7 male and female embryos.

All values are plotted in Log_2 scale. Black lines show H3K9me3 in 10Kb windows across the genome. Regions enriched for H3K9me3 are shown in red/dark orange.

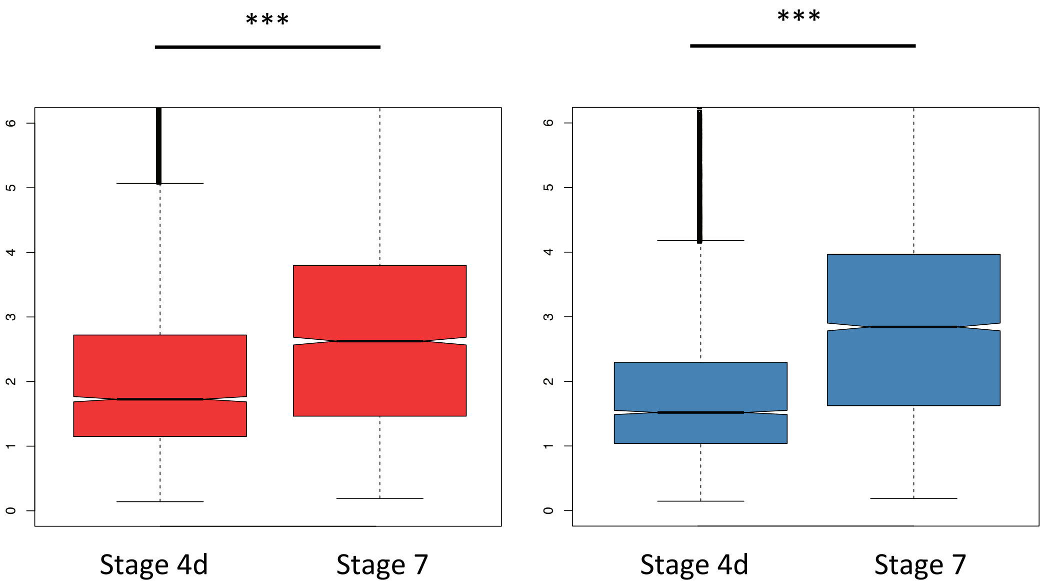


Figure 3. Global increase in heterochromatin in pericentromeric regions in females (red) and males (blue) from stage 4d to stage 7.

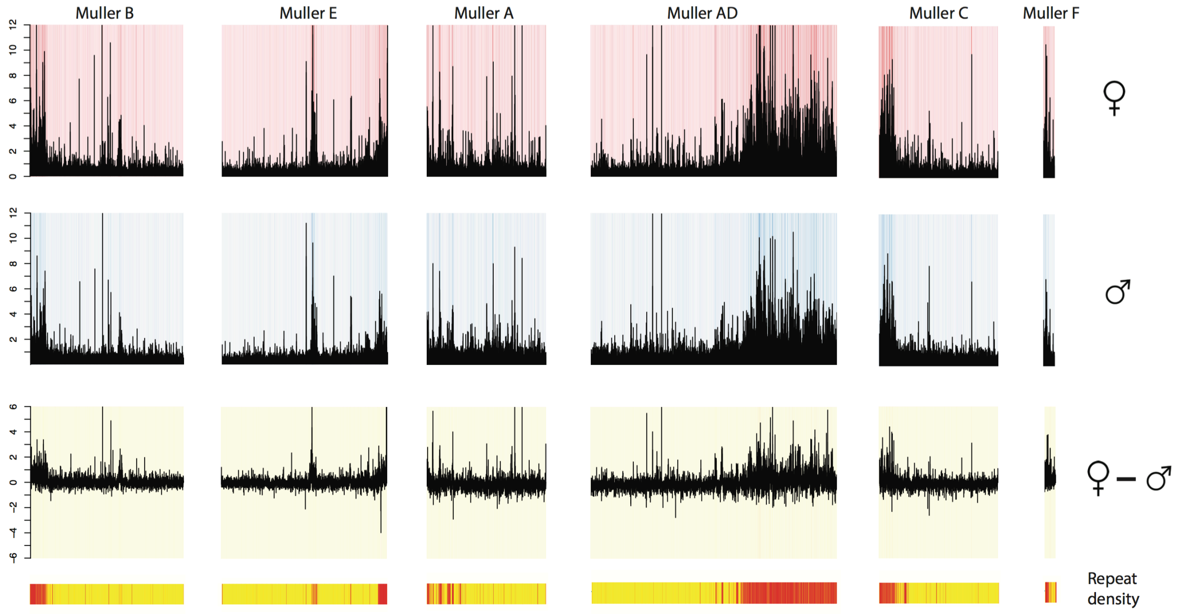


Figure 4. Genome wide H3K9me3 enrichment plots for stage 4d females (in red), males (in blue) and subtraction plot of female – male H3K9me3 enrichment (in light yellow).

The bars at the bottom of the plot show the repeat density along the length of the chromosomes (Regions with high repeat density are shown in red and regions with low repeat density are shown in yellow).

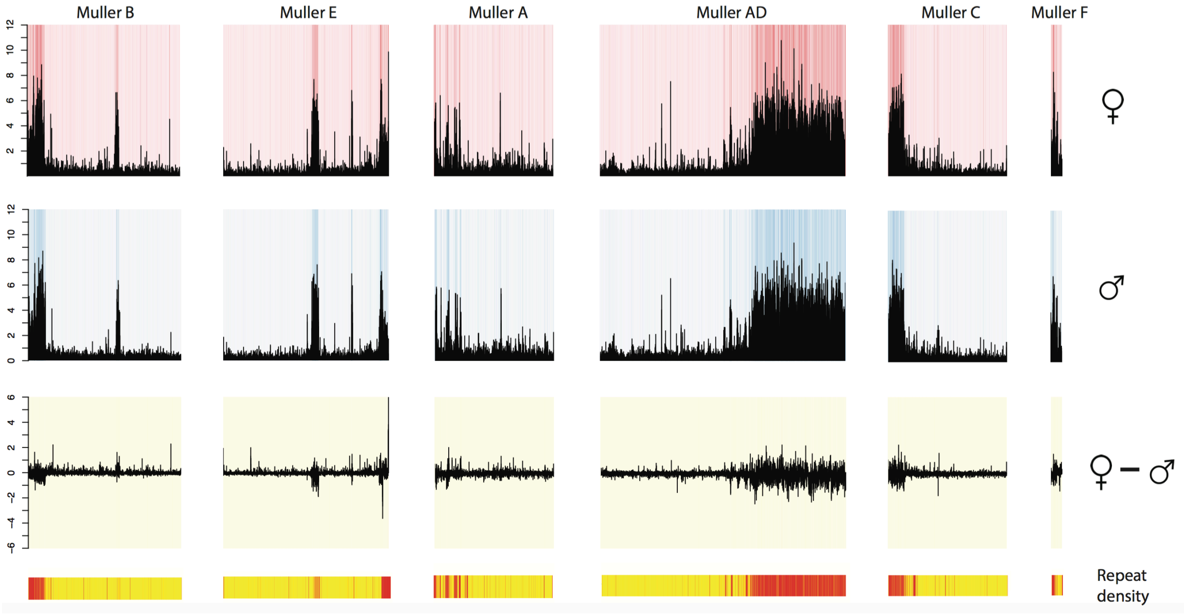


Figure 5. Genome wide H3K9me3 enrichment plots for stage 7 females (in red), males (in blue) and subtraction plot of female – male H3K9me3 enrichment (in light yellow).

The bars at the bottom of the plot show the repeat density along the length of the chromosomes (Regions with high repeat density are shown in red and regions with low repeat density are shown in yellow).

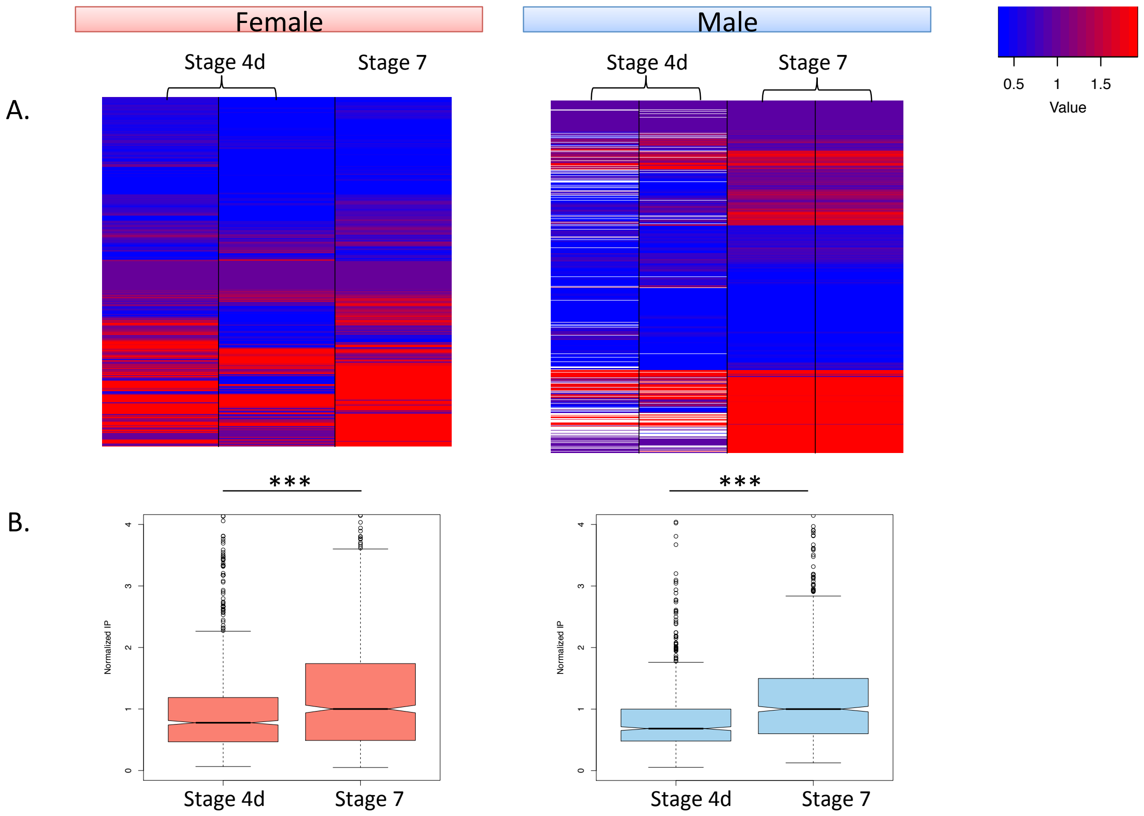


Figure 6. H3K9me3 enrichment at repeats in stage 4d and stage 7 males and females.

A. H3K9me3 enrichment at repeats in female and male stage 4d and stage 7 embryos. B. Boxplots showing a significant increase in H3K9me3 binding at repeats from stage 4d to stage 7 in both males and females.

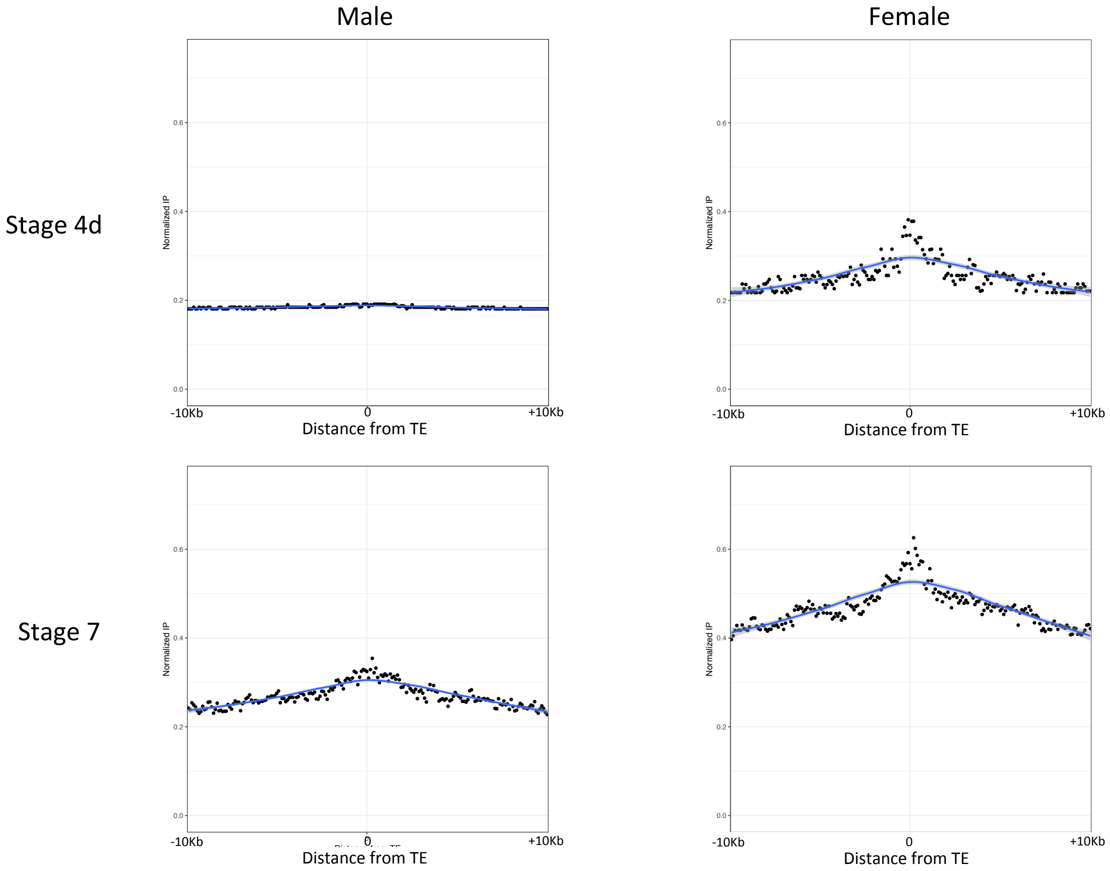


Figure 7. Metagene plots showing H3K9me3 enrichment in flanking regions of 1743 euchromatic TE insertions greater than 1Kb in size.

H3K9me3 enrichment in 10Kb flanking regions were plotted on either side of the insertion (100bp window size).

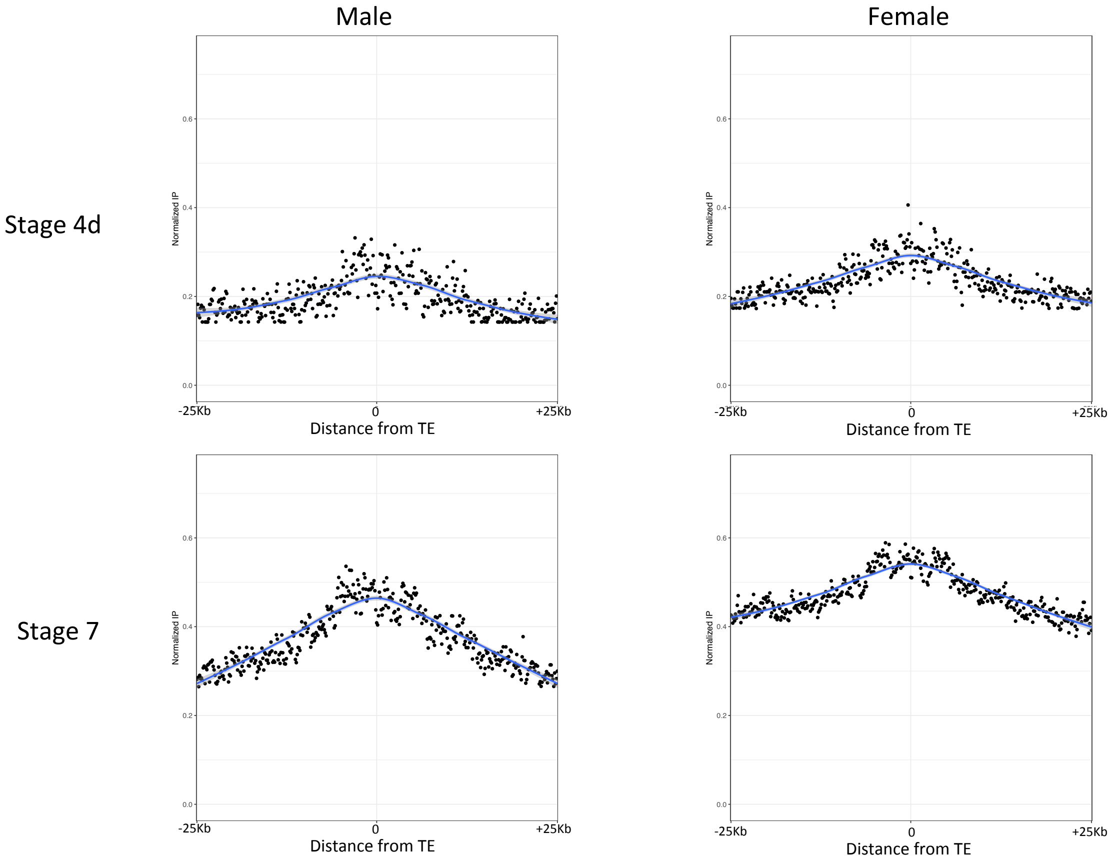


Figure 8. **Metagene plots showing H3K9me3 enrichment in 25Kb flanking regions of 708 euchromatic TE insertions greater than 1Kb in size that lie in regions of the genome with large number of piRNA reads mapping to them.**

H3K9me3 enrichment in 25Kb flanking regions were plotted on either side of the insertion (125bp window size).

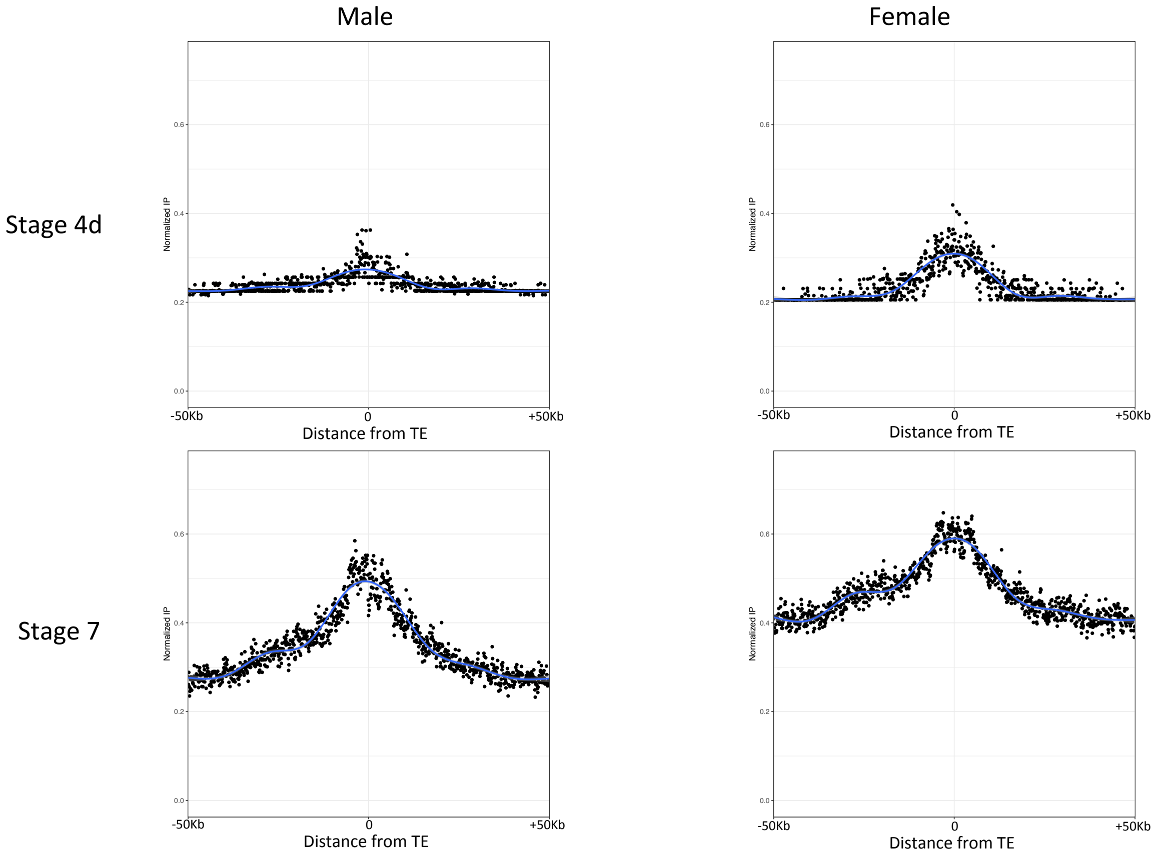


Figure 9. **Metagene plots showing H3K9me3 enrichment in 50Kb flanking regions of 681 euchromatic TE insertions greater than 1Kb in size that lie in regions of the genome with large number of piRNA reads mapping to them.**

H3K9me3 enrichment in 50Kb flanking regions were plotted on either side of the insertion (100bp window size)

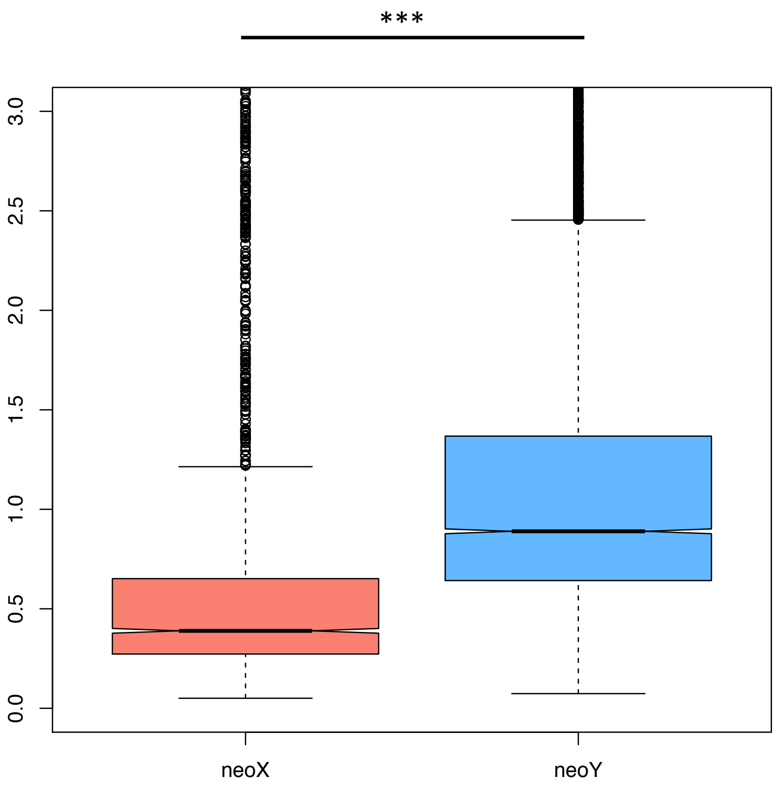


Figure 10. H3K9me3 enrichment on the neoX versus neoY in stage 7 male embryos.

H3K9me3 enrichment is significantly higher on the neoY chromosome compared to the neoX chromosome.

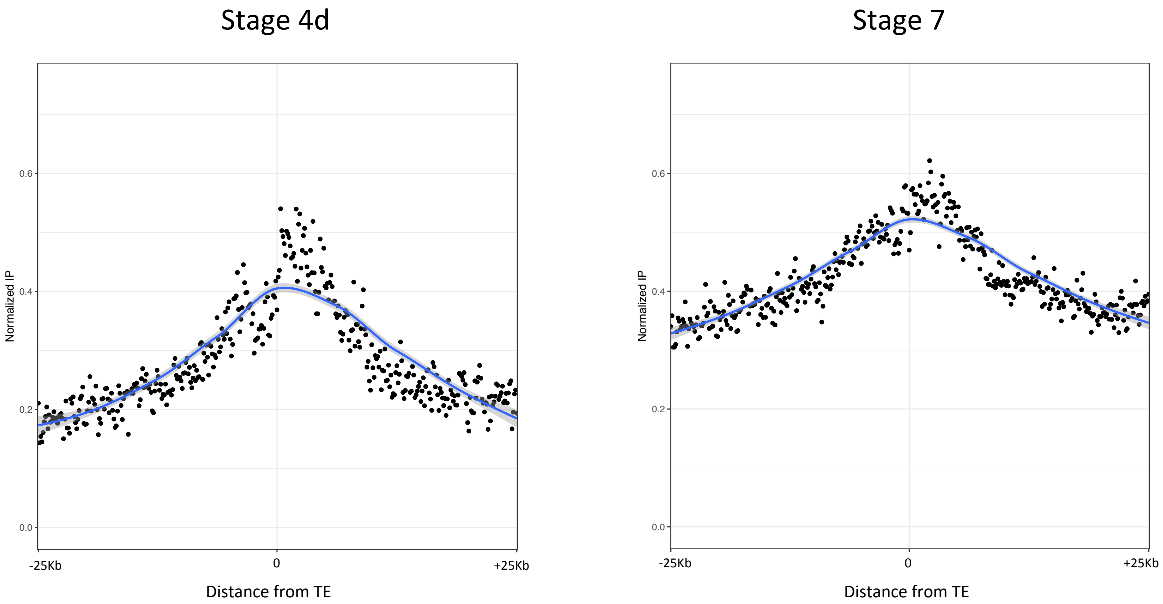
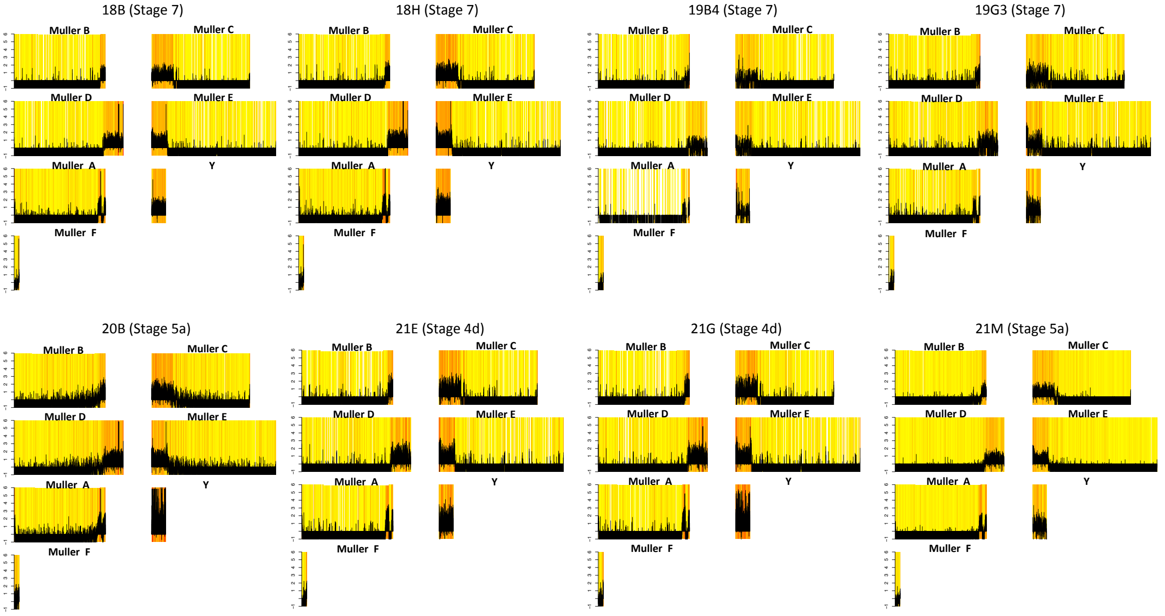


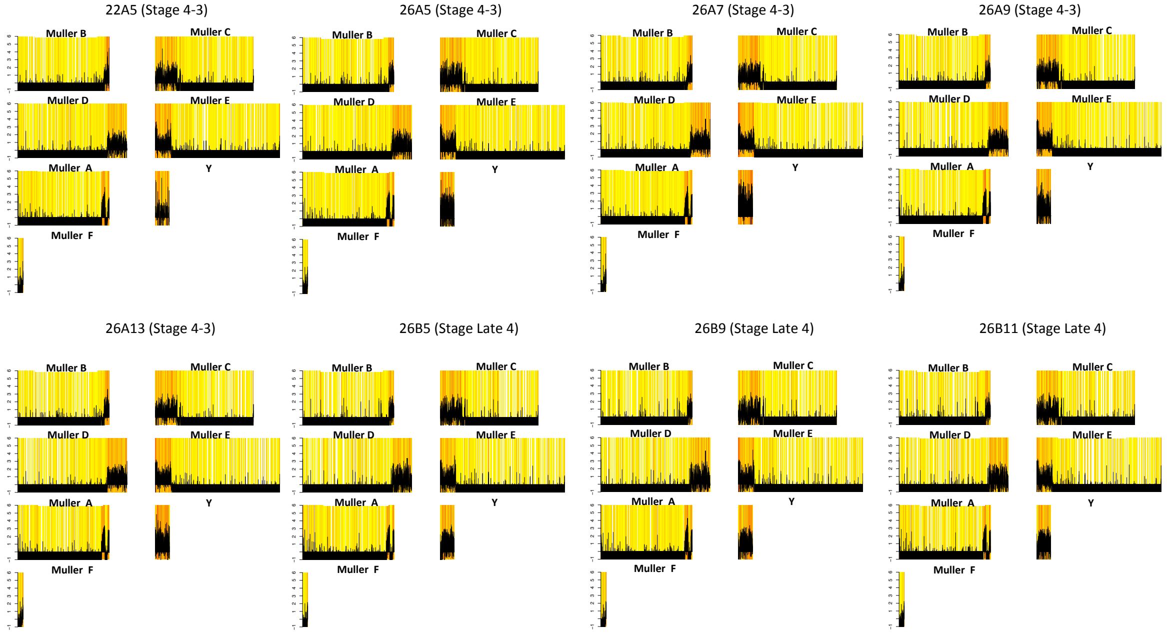
Figure 11. Metagene plots showing H3K9me3 enrichment in 25 Kb flanking regions of 460 Repeat insertions greater than 1Kb in size that lie in regions of the neoY chromosome with the highest number of piRNA reads mapping to them (Top 200 genomic 10Kb windows).

H3K9me3 enrichment in 25Kb flanking regions were plotted on either side of the insertion (100bp window size).

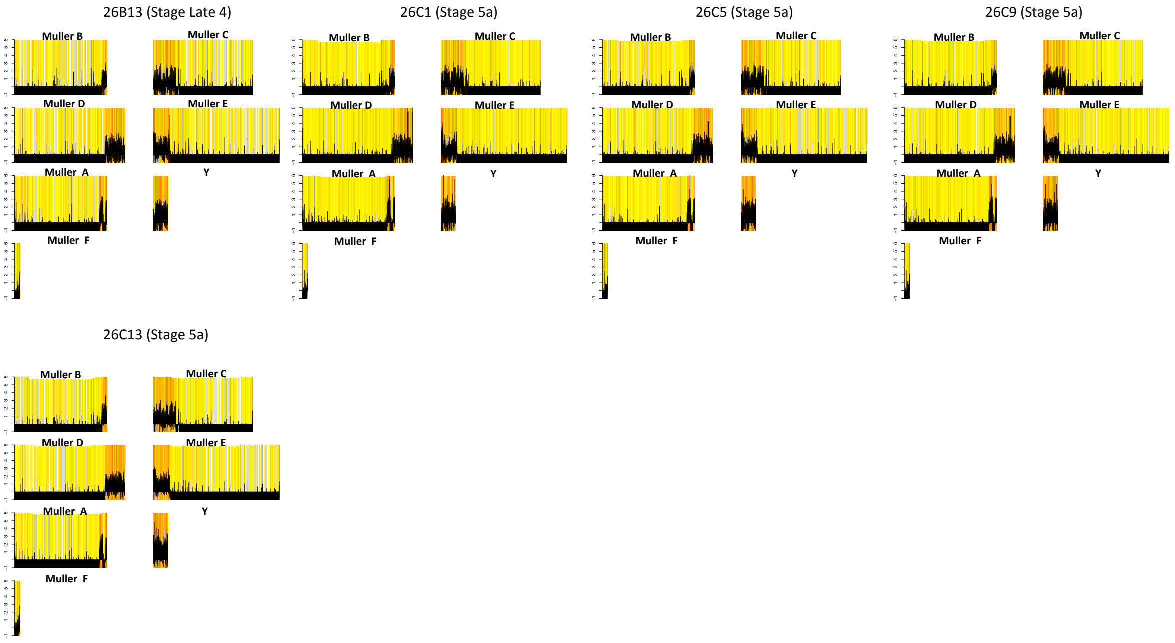
Supplementary Figures



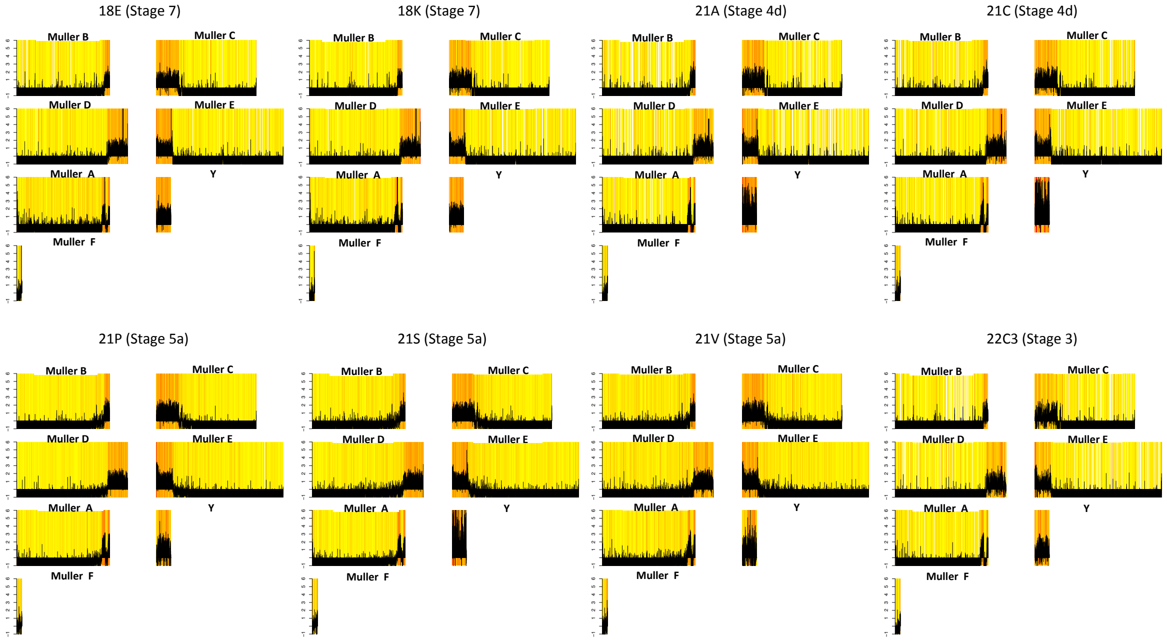
Supplementary figure 1a. H3K9me3 binding profiles for spikes used for female *D.miranda* embryo samples.



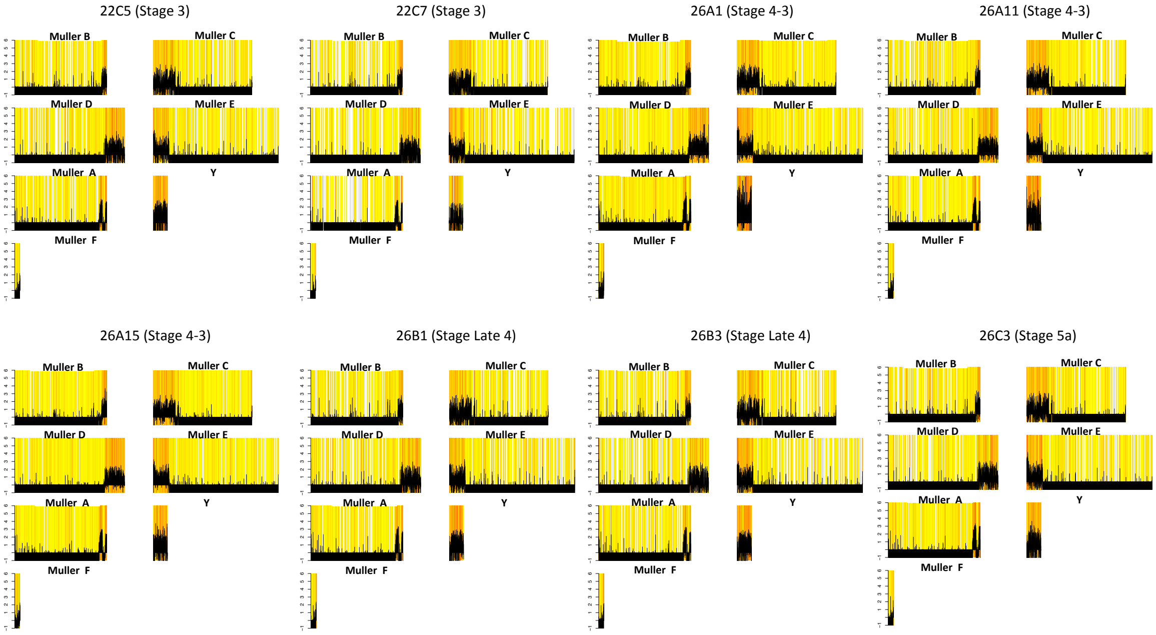
Supplementary figure 1b. H3K9me3 binding profiles for spikes used for female *D.miranda* embryo samples.



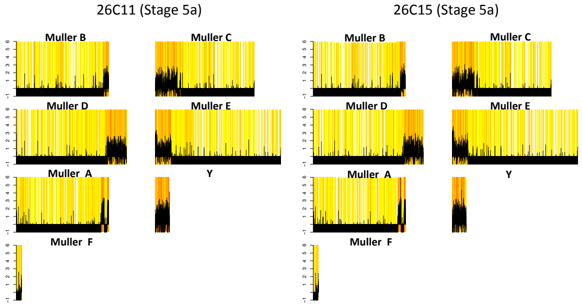
Supplementary figure 1c. H3K9me3 binding profiles for spikes used for female *D.miranda* embryo samples.



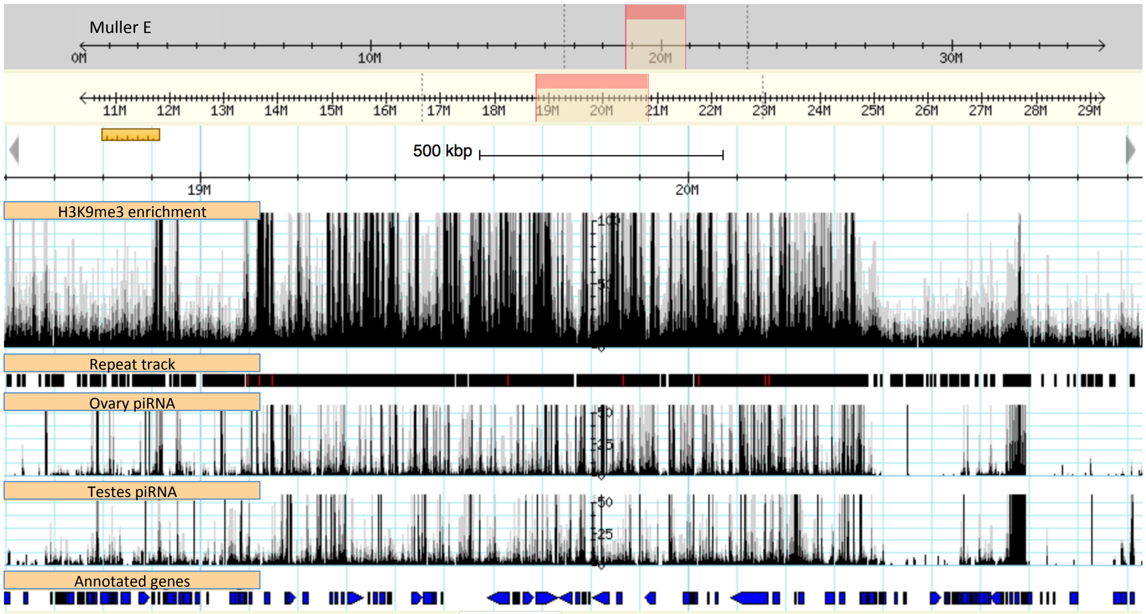
Supplementary figure 1d. H3K9me3 binding profiles for spikes used for male *D.miranda* embryo samples.



Supplementary figure 1e. H3K9me3 binding profiles for spikes used for male *D.miranda* embryo samples.

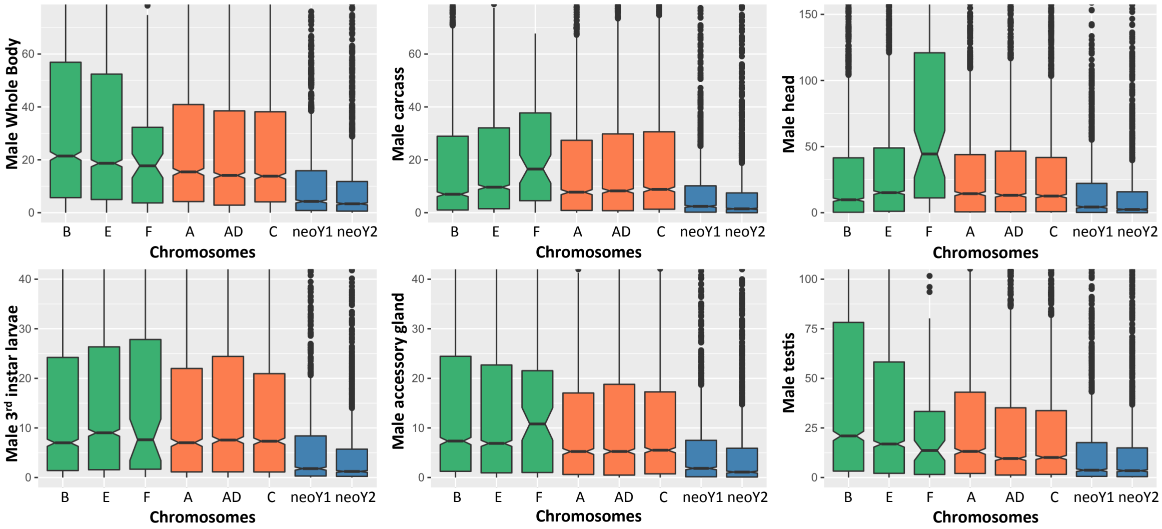


Supplementary figure 1f. H3K9me3 binding profiles for spikes used for male *D.miranda* embryo samples.

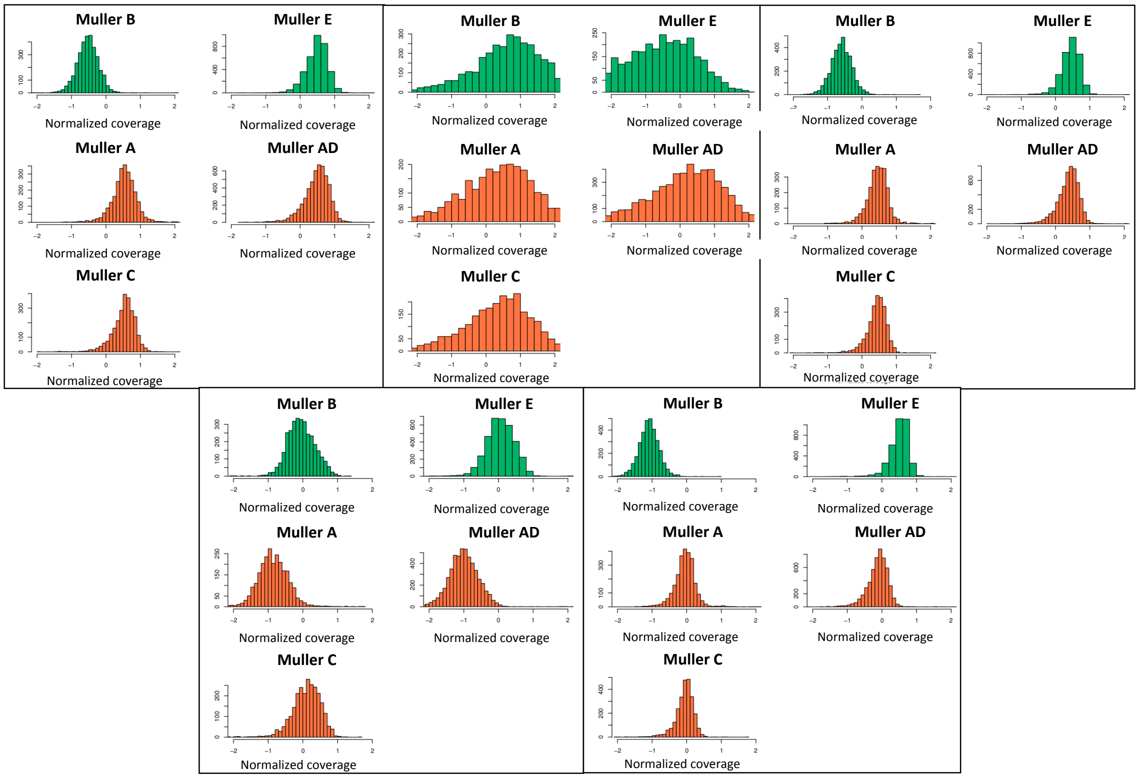


Supplementary Figure 2. Heterochromatin island on Muller E.

A large (~2Mb) heterochromatin island on Muller E, that has large number of ovary and testes piRNA reads mapping.



Supplementary Figure 3. Box plots showing gene expression (in TPM) in different tissues. NeoY has a significantly lower expression (p-value Wilcox<10-5) compared to all other chromosomes.



Supplementary Figure 4. Five Aneuploid embryos. Normalized coverage histograms for autosomes in green and X-linked chromosomes in red.

Supplementary Tables

Supplementary Table 1. Embryo samples collected for study (alignment statistics for PE sequencing reads)

MALE embryo samples

		CHIP						INPUT					
		Number of reads aligned			Percentage reads aligned			Number of reads aligned			Percentage reads aligned		
Sample	Stage	Mir sample	Mel sample	Total sample	%Mir sample	%Mel sample	Total% sample	Mir input	Mel input	Total input	%Mir input	%Mel input	Total% input
18E	7	3.6E+07	1.7E+07	5.4E+07	64.3	30.47	94.77	3.5E+07	1.7E+07	5.2E+07	65.49	31.93	97.42
18K	7	3.6E+07	8028970	4.4E+07	80.48	18.2	98.68	3.7E+07	8689777	4.6E+07	79.62	18.59	98.21
21A	4d	4166584	2.3E+07	2.7E+07	12.76	70.25	83.01	1.5E+07	2.3E+07	3.8E+07	38.1	57.71	95.81
21C	4d	7537896	1.2E+07	1.9E+07	26.98	42.14	69.12	1.4E+07	6192679	2E+07	66.11	28.9	95.01
21P	5a	1.7E+07	8568363	2.5E+07	55.21	28.58	83.79	3.5E+07	8944634	4.4E+07	76.35	19.75	96.1
21S	5a	1.7E+07	4928183	2.2E+07	56.17	16.1	72.27	2.8E+07	2605266	3.1E+07	88.01	8.2	96.21
21V	5a	1534822	825972	2360794	7.92	4.26	12.18	1.5E+07	3455005	1.8E+07	75.87	17.95	93.82
22C1	3	95606	1311727	1407333	1.98	27.23	29.21	8706242	1.2E+07	2.1E+07	38.94	53.81	92.75
22C3	3	795877	795877	1591754	2.53	42.31	44.84	2793963	9740666	1.3E+07	20.65	72	92.65
22C5	3	1436522	1436522	2873044	5.25	29.97	35.22	7849001	6059330	1.4E+07	53.78	41.52	95.3
22C7	3	1432643	1432643	2865286	7.93	42.55	50.48	7409315	5767022	1.3E+07	53.24	41.44	94.68
26A11	4-3	2043322	2043322	4086644	16.42	51.8	68.22	6365099	5675572	1.2E+07	50.04	44.61	94.65
26A15	4-3	3754639	9610971	1.3E+07	19.51	49.93	69.44	6366934	6688518	1.3E+07	47.74	50.15	97.89
26A1	4-3	638927	4657605	5296532	4.72	34.44	39.16	4543837	6789095	1.1E+07	39.07	58.37	97.44
26B1	late4	2544521	7291959	9836480	18.22	52.2	70.42	7335222	5618661	1.3E+07	55.41	42.44	97.85
26B3	late4	1736264	1736264	3472528	9.91	66.43	76.34	7317928	7593844	1.5E+07	47.62	49.41	97.03
26C11	5a	2325815	2325815	4651630	25.16	32.08	57.24	9382936	3138581	1.3E+07	73.6	24.62	98.22
26C15	5a	3513455	3619642	7133097	24.13	24.86	48.99	1.5E+07	4059380	1.9E+07	76.85	20.89	97.74
26C3	5a	2651374	2415929	5067303	26.33	24	50.33	8110815	2339019	1E+07	76.26	21.99	98.25

Supplementary Table 1. (contd.) Embryo samples collected for study (alignment statistics for PE sequencing reads)

FEMALE embryo samples

		CHIP						INPUT					
		Number of reads aligned			Percentage reads aligned			Number of reads aligned			Percentage reads aligned		
Sample	Stage	Mir sample	Mel sample	Total sample	%Mir sample	%Mel sample	Total% sample	Mir input	Mel input	Total input	%Mir input	%Mel input	Total% input
18B	7	4E+07	1.1E+07	5.1E+07	76.72	21.12	97.84	3.5E+07	1.1E+07	4.6E+07	74.48	22.81	97.29
18H	7	3.5E+07	1.3E+07	4.8E+07	72.28	25.53	97.81	3.5E+07	9961323	4.5E+07	76.37	21.95	98.32
19B4	7	2.2E+07	4537251	2.7E+07	75.96	15.64	91.6	1.9E+07	4387196	2.3E+07	80	18.83	98.83
19G3	7	1.4E+07	1E+07	2.4E+07	49.83	36.86	86.69	2E+07	5730184	2.6E+07	77.15	21.98	99.13
20B	5a	2863175	1212835	4076010	55.78	23.63	79.41	2.7E+07	6429455	3.3E+07	78.73	18.92	97.65
21E	4d	4494295	1.7E+07	2.1E+07	16.23	59.6	75.83	9278166	9443726	1.9E+07	47	47.84	94.84
21G	4d	5188017	1.5E+07	2E+07	19.66	55.17	74.83	1.2E+07	9787592	2.2E+07	53.83	42.2	96.03
21M	5a	5026967	3129069	8156036	17.68	11	28.68	4.5E+07	1.1E+07	5.6E+07	77.59	19.34	96.93
22A1	4-3	7745930	9205923	1.7E+07	25.1	29.83	54.93	523103	181419	704522	72.29	25.07	97.36
22A5	4-3	5680891	9099996	1.5E+07	22.4	35.89	58.29	8317589	5587963	1.4E+07	57.45	38.59	96.04
26A13	4-3	1047297	7081411	8128708	7.07	47.83	54.9	7260104	5181410	1.2E+07	56.24	40.14	96.38
26C9	5a	3376975	4114012	7490987	30.82	37.55	68.37	7504179	3107569	1.1E+07	69.39	28.73	98.12
26A5	4-3	2154836	5849658	8004494	18.09	49.12	67.21	6370315	4965993	1.1E+07	54.48	42.47	96.95
26A7	4-3	2531268	6804994	9336262	20.46	55	75.46	7948612	5343793	1.3E+07	58.11	39.06	97.17
26A9	4-3	2447874	5453511	7901385	19.34	43.09	62.43	7198594	5901787	1.3E+07	52.73	43.23	95.96
26B11	Late 4	2708940	8979804	1.2E+07	15.36	50.9	66.26	1.1E+07	7151537	1.8E+07	57.92	39.28	97.2
26B13	Late 4	924137	5748583	6672720	8.65	53.8	62.45	7305847	4449640	1.2E+07	60.52	36.86	97.38
26B5	Late 4	1561129	4987801	6548930	9.84	31.45	41.29	7373600	7559217	1.5E+07	47.52	48.71	96.23
26B9	Late 4	3067841	6754878	9822719	23.84	52.48	76.32	8405497	5982812	1.4E+07	56.55	40.25	96.8
26C13	5a	2917840	6522061	9439901	22.6	50.52	73.12	1.1E+07	4826467	1.6E+07	67.85	30.17	98.02
26C1	5a	2795848	4244739	7040587	26.93	40.89	67.82	6481767	3864107	1E+07	54.62	32.56	87.18
26C5	5a	4386225	4678065	9064290	35.75	38.13	73.88	7278957	2862939	1E+07	69.1	27.18	96.28