Submission for Article
Fit in the Discoveries

Gene-level, but not chromosome-wide, divergence between a very young house fly proto-Y chromosome and its homologous proto-X chromosome

Jae Hak Son[1, 2, 3] and Richard P. Meisel[1, 4]

1. Department of Biology and Biochemistry, University of Houston, Houston, TX, USA

2. Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

3. Jae Hak Son: jaehak.son.225@gmail.com

4. Richard P. Meisel: rpmeisel@uh.edu

Abstract

X and Y chromosomes are usually derived from a pair of homologous autosomes, which then diverge from each other over time. Although Y-specific features have been characterized in sex chromosomes of various ages, the earliest stages of Y chromosome evolution remain elusive. In particular, we do not know whether early stages of Y chromosome evolution consist of changes to individual genes or happen via chromosome-scale divergence from the X. To address this question, we quantified divergence between young proto-X and proto-Y chromosomes in the house fly, *Musca domestica*. We compared proto-sex chromosome sequence and gene expression between genotypic (XY) and sex-reversed (XX) males. We find evidence for sequence divergence between genes on the proto-X and proto-Y, including five genes with mitochondrial functions. There is also an excess of genes with divergent expression between the proto-X and proto-Y, but the number of genes is small. This suggests that individual proto-Y genes, but not the entire proto-Y chromosome, have diverged from the proto-X. We identified one gene, encoding an axonemal dynein assembly factor (which functions in sperm motility), that has higher expression in XY males than XX males because of a disproportionate contribution of the proto-Y allele to gene expression. The up-regulation of the proto-Y allele may be favored in males because of this gene's function in spermatogenesis. The evolutionary divergence between proto-X and proto-Y copies of this gene, as well as the mitochondrial genes, is consistent with selection in males affecting the evolution of individual genes during early Y chromosome evolution.

Introduction

In many organisms with two separate sexes, a gene on a sex chromosome determines whether an individual develops into a male or female. In XX/XY sex chromosome systems, males are the heterogametic sex (XY genotype), and females are homogametic with the XX genotype (Bull 1983). Most X and Y chromosomes are derived from a pair of ancestral autosomes. For example, one copy of the autosome can obtain a male-determining gene and become a proto-Y chromosome, and the homologous chromosome without the male-determiner becomes a proto-X. As the proto-X and proto-Y chromosomes diverge from each other over time, they become differentiated X and Y chromosomes (Bull 1983; Charlesworth et al. 2005). Sex chromosomes have originated and diverged from each other in multiple independent evolutionary lineages (Bachtrog et al. 2014; Beukeboom and Perrin 2014).

Despite their independent origins, non-homologous Y chromosomes share many common features across species (Charlesworth et al. 2005). First, "masculinization" occurs because male-limited inheritance of the Y chromosome favors the fixation of male-beneficial genetic variants (Rice 1996a). Second, suppressed recombination between the X and Y chromosomes evolves, possibly due to sexually antagonistic selection, meiotic drive, or genetic drift (Charlesworth 2017; Charlesworth 2018; Ponnikas et al. 2018). Third, "degeneration" occurs in nonrecombining regions—functional genes that were present on ancestral autosomes become pseudogenes on the Y chromosome because suppressed recombination between the X and Y inhibits the purging of deleterious mutations in Y-linked genes (Muller's ratchet) and enhances the effects of hitchhiking (Charlesworth and Charlesworth 2000; Bachtrog 2013; Vicoso 2019). Other common features of Y chromosomes are repetitive sequences and enlarged heterochromatic regions due to reduced efficacy of purifying selection caused by suppressed recombination and a small effective population size (Skaletsky et al. 2003). In some cases, a mechanism evolves to compensate for the haploid dosage of X-linked genes in males, but this is not always the case (Mank 2013; Gu and Walters 2017).

Many features of Y chromosomes are thought to emerge shortly after an autosome obtains a new male-determining locus or becomes Y-linked. For example, recombination suppression has been

66 considered to come after the emergence of a new sex-determining locus on a proto-Y
67 chromosome to favor the co-inheritance of the sex-determining locus and male-
68 beneficial/female-detrimental sexually antagonistic alleles (Orzack et al. 1980; van Doorn and
69 Kirkpatrick 2007; Roberts et al. 2009; van Doorn and Kirkpatrick 2010). Additional sexually
70 antagonistic alleles on the proto-Y chromosome are predicted to trigger progressive spread of the
71 nonrecombining region along the chromosome (Rice 1987; van Doorn and Kirkpatrick 2007).
72 Although these features have been characterized in sex chromosomes of various ages and
73 degeneration levels (Bachtrog 2013; Zhou et al. 2014), the very first stages of Y chromosome
74 evolution are poorly understood because of a lack of extremely young sex chromosome systems.
75 Recent studies of young sex chromosomes have identified multiple types of X-Y differentiation,
76 including suppressed recombination, Y chromosome gene loss, and X chromosome dosage
77 compensation (Bergero et al. 2013; Mahajan et al. 2018; Darolti et al. 2019; Krasovec et al.
78 2019), which makes it challenging to determine which type of differentiation occurs first.
79
80 The extent to which the early evolution of sex chromosomes is dominated by chromosome-wide
81 X-Y divergence versus changes in individual genes remains unclear. This study addresses that
82 shortcoming by determining how a young proto-Y chromosome has differentiated from its
83 homologous proto-X chromosome shorty after its emergence. We are especially interested in
84 how gene expression differences accumulate between the proto-Y and proto-X chromosomes. As
85 the proto-Y and proto-X chromosomes diverge, it is expected that alleles on the proto-Y
86 chromosome are up- or down-regulated because of *cis*-regulatory sequence differences that
87 contribute to proto-Y gene expression (Zhou and Bachtrog 2012a; Zhou and Bachtrog 2012b;
88 Wei and Bachtrog 2019). These *cis*-regulatory effects may be especially important for the
89 expression of sexually antagonistic (male-beneficial/female-deleterious) alleles and degeneration
90 of functional genes (Rice 1984; Zhou and Bachtrog 2012a). Degeneration of Y-linked genes has
91 been shown to be accompanied by decreased expression as a result of relaxed selective
92 constraints (Zhou and Bachtrog 2012a; Wei and Bachtrog 2019). However, the accumulation of
93 gene expression differences separately from degeneration during the very earliest stages of sex
94 chromosome evolution are not well understood.
95

96 We used the house fly, *Musca domestica*, as a model system to study the early evolution of sex
97 chromosomes because it has very young proto-sex chromosomes that are still segregating as
98 polymorphisms within natural populations (Hamm et al. 2015). The *M. domestica male*
99 *determiner* (*Mdmd*) can be found on what was historically called the Y chromosome (Y^M) and on
100 at least three other chromosomes (Sharma et al. 2017). The house fly Y^M (and X) chromosome
101 has fewer than 100 genes, while the other five chromosomes each have >2,000 genes (Meisel
102 and Scott 2018). Each chromosome carrying *Mdmd*, including Y^M, is a recently derived proto-Y
103 chromosome (Meisel et al. 2017). *Mdmd* arose in the house fly genome after divergence from
104 stable fly (*Stomoxys calcitrans*) and horn fly (*Haematobia irritans*), within the past 27 million
105 years (Sharma et al. 2017; Meisel et al. 2020). This provides an upper-bound on the age of the
106 house fly proto-Y chromosomes, although the minimal sequence and morphological divergence
107 between the proto-Y and proto-X chromosomes (Boyes et al. 1964; Hediger et al. 1998; Meisel
108 et al. 2017) suggest they are much younger than that. It is not clear the extent to which the house
109 fly proto-Y chromosomes are masculinized or degenerated. A previous study revealed a small,
110 but significant, effect of the proto-Y chromosomes on gene expression (Son et al. 2019).
111 However, it could not resolve if the expression differences are the result of changes in the
112 expression of the proto-Y copies, proto-X copies, or both.
113
114 In this study, we tested if one house fly proto-Y chromosome, the third chromosome carrying
115 *Mdmd* (III^M), has evidence of gene-by-gene or chromosome-wide differentiation from its
116 homologous proto-X chromosome by evaluating DNA sequence and gene expression differences
117 between proto-Y genes and their proto-X counterparts. We selected this proto-sex chromosome
118 (as opposed to other house fly proto-Y chromosomes) for three reasons. First, III^M is one of the
119 two most common proto-Y chromosomes found in natural populations (Hamm et al. 2015).
120 Second, the other common proto-Y (known as Y^M) has fewer than 100 genes, which is over one
121 order of magnitude less than III^M (Meisel et al. 2017; Meisel and Scott 2018). More genes on the
122 third chromosome gives us greater power to detect divergence between the proto-Y and proto-X.
123 Third, we are able to create sex-reversed males (genotypic females that are phenotypically male)
124 with the same genetic background as III^M males (Hediger et al. 2010). This allows us to compare
125 gene expression between phenotypic males that only differ in whether or not they carry the III^M
126 proto-Y chromosome (Son et al. 2019).

127

128 Our expectation for the extent of gene-by-gene versus chromosome-wide differentiation between
129 the proto-X and proto-Y chromosomes will depend on the extent of X-Y recombination. We
130 expect gene-by-gene differentiation on young sex chromosomes if there is free recombination
131 that prevents the effects of Muller's ratchet and genetic hitchhiking (Charlesworth and
132 Charlesworth 2000). Recombination between the house fly proto-X and proto-Y is possible
133 because, unlike *Drosophila*, there may be recombination in male house flies (Feldmeyer et al.
134 2010). In addition, female house flies can carry a proto-Y chromosome if they have a female-
135 determining allele on another chromosome (Mcdonald et al. 1978; Hediger et al. 2010; Hamm et
136 al. 2015), which would provide additional opportunities for recombination. However, if there are
137 recombination suppressors (such as chromosomal inversions that differentiate the proto-Y and
138 proto-X), this would promote chromosome-wide divergence between the proto-X and proto-Y
139 via Muller's ratchet and hitchhiking (Ponnikas et al. 2018).

140

141

142 **Results and Discussion**

143

144 **DNA sequence divergence between the proto-Y and proto-X chromosomes**

145

146 We used RNA-seq data to identify single nucleotide polymorphisms (SNPs) and small
147 insertions/deletions (indels) within genes in genotypic (III^M/III) and sex-reversed (III/III) male
148 house flies (Son et al. 2019). For each gene, we counted the number of sites that are
149 heterozygous in either genotypic or sex-reversed males (i.e., two alleles in at least one genotype).
150 We then calculated the percent of those sites (per gene) that are heterozygous only in the
151 genotypic males. This value is 0% if heterozygous sites are only in sex-reversed males, it is
152 100% if heterozygous sites are only in genotypic males, and it is 50% if the same number of
153 heterozygous sites are found in both genotypes. We found that the genotypic males have an
154 excess of heterozygous sites in third chromosome genes, relative to the sex-reversed males
155 (Figure 1; $P < 10^{-16}$ in a Wilcoxon rank sum test comparing percent heterozygous sites in genes
156 on the third chromosome with genes on the other chromosomes). This is consistent with a
157 previous comparison between III^M males and Y^M males (Meisel et al. 2017), and it suggests that

158 the sequences of genes on the III^M proto-Y chromosome are differentiated from the copies on the
159 proto-X (i.e., the standard third chromosome).
160
161 We expect that the ancestral X chromosome would have the same levels of heterozygosity in
162 genotypic males (X/X; III^M/III) and sex-reversed males (X/X; III/III) due to the presence of two
163 copies of the X chromosome in both genotypes. However, the III^M males have elevated
164 heterozygosity on the X chromosome (Figure 1; $P = 8.32\times10^{-13}$ in a Wilcoxon rank sum test
165 comparing the X chromosome with chromosomes I, II, IV, and V). This was also observed in a
166 comparison between III^M males and Y^M males (Meisel et al. 2017), and the cause of the elevated
167 X chromosome heterozygosity in III^M males remains unresolved. One possible explanation for
168 elevated X chromosome heterozygosity is that the III^M chromosome was created by a fusion
169 between the third chromosome and the Y^M chromosome. Because Y^M has nearly identical gene
170 content as the X chromosome (Meisel et al. 2017), this would cause III^M males to have three
171 copies of X chromosome genes, increasing the likelihood that III^M males are heterozygous at any
172 given site on the X chromosome. This hypothesis remains to be tested.
173
174 We further tested for divergence between the III^M proto-Y chromosome and its homologous
175 proto-X by assembling a III^M male genome from the same strain as our RNA-seq data using
176 Oxford Nanopore reads. We were specifically interested in identifying contigs that were
177 separately assembled from homologous regions on the proto-X and proto-Y chromosomes (i.e.,
178 gametologs; Garcia-Moreno and Mindell 2000). Assembly into separate contigs would provide
179 evidence for divergence between the proto-X and proto-Y sequences. Our III^M male assembly is
180 smaller (427 Mb) than the reference genome assembly (691 Mb; Scott et al. 2014) and flow
181 cytometry estimates of genome size (~1,000 Mb; Picard et al. 2012), suggesting that we are
182 missing 30-50% of the genome sequence in our assembly. The reduced assembled size of our
183 III^M male genome is likely the result of low sequencing coverage. Nonetheless, we should be
184 able to identify X-Y divergence, albeit with reduced power relative to a more complete genome
185 assembly. We took two approaches to identify separate proto-X and proto-Y contigs in our
186 assembly: 1) identifying genes on the same contig as the male-determining gene (*Mdmd*); and
187 2) testing for contigs containing sequences that are enriched in males relative to females
188 (Carvalho and Clark 2013). Our approaches will identify regions on the proto-sex chromosome

189 that contain genes that can be assigned to the third chromosome, are present on both the proto-X
190 and proto-Y with sufficient divergence to assemble separately, and with both the proto-X and
191 proto-Y gametologs assembled in our III^M male genome. We only identified two loci, one from
192 each of our two approaches, with sufficient X-Y divergence to assemble into separate proto-X
193 and proto-Y contigs.
194
195 We identified one contig containing *Mdmd* in the III^M assembly (ctg2382) that contains the same
196 three genes as another contig (ctg1607) without *Mdmd* (Figure 2A). We infer the contig with
197 *Mdmd* to be the proto-Y sequence, and the one without to be from the proto-X. There are two
198 additional contigs containing *Mdmd* in our III^M genome assembly, but we did not find any
199 corresponding proto-X contigs for these (Supplementary Figure 1). These other two contigs with
200 *Mdmd* either contain loci without sufficient X-Y divergence to assemble into separate proto-X
201 and proto-Y contigs, or we did not sequence deep enough to capture enough proto-X reads to
202 assemble the proto-X gametolog. The male-determining regions of the house fly proto-Y
203 chromosomes contain both complete and truncated copies of *Mdmd* (Sharma et al. 2017), and
204 ctg2382 contains two truncated copies of *Mdmd*. The genomic region corresponding to this
205 contig is found in a single scaffold of the reference genome (NW_004774683.1), which was
206 previously assigned to the third chromosome (Meisel and Scott 2018). There are three genes at
207 this locus, and they are present in the same order in the reference genome, the proto-Y contig,
208 and proto-X contig (Figure 2A).
209
210 We found the second locus using a *k*-mer comparison approach to test for contigs containing
211 sequences that are enriched in males relative to females (Carvalho and Clark 2013). Here, we
212 compared *k*-mers in our Nanopore III^M male reads and the Illumina reads used to generate the
213 reference (female) genome assembly (Scott et al. 2014) to our III^M male assembly. We then
214 calculated the percent of each contig in our III^M assembly that is unmatched by female reads
215 (%UFR), which we expect to be 100% for a Y chromosome contig that only contains male-
216 specific *k*-mers. We did not identify any contigs with greater than 27.4%UFR (Supplementary
217 Figure 1), which is consistent with the minimal sequence divergence between the proto-Y and
218 proto-X chromosomes (Meisel et al. 2017). To determine an expectation for what constitutes
219 a %UFR of a proto-Y contig in our assembly, we calculated %UFR for the three contigs that

220 each contain copies of *Mdmd* (Supplementary Figure 1). The largest %UFR of any contig with
221 *Mdmd* is 13.8% (ctg3705). We used this value as a threshold to find proto-Y contigs that do not
222 contain *Mdmd*.
223
224 We found 275 contigs with a %UFR>13.8, nine of which (Supplementary Table 1) contain
225 protein-coding genes that were previously assigned to the third chromosome (Meisel and Scott
226 2018). One gene (LOC101898200, which encodes histone H3) was found on six different contigs
227 with %UFR>13.8 (and an additional 27 contigs with %UFR<13.8). Histone gene clusters are a
228 common feature of most metazoan genomes, and the multiple copies are under tight regulatory
229 control (McKay et al. 2015; Duronio and Marzluff 2017).These contigs with high %UFR and
230 copies of histone H3 genes are either false positives or not of substantial phenotypic relevance.
231 Two of the other contigs with %UFR>13.8 (ctg3539 and ctg8407) each contain 1-2 genes that
232 are found as multiple copies on other contigs (all with a %UFR<13.8). One of those genes
233 (LOC109612838 on ctg3539) encodes a protein with predicted reverse transcriptase and RNase
234 H activity, suggesting that it is a retrotransposon (Finnegan 2012). Another gene
235 (LOC109613819 on ctg8407) encodes a protein with zinc finger and integrase domains, also
236 consistent with a transposable element (Volff 2006). We exclude these contigs from our
237 subsequent analysis because they likely reflect transposable element expansions within the house
238 fly genome. We only found one contig with a %UFR>13.8 (ctg2522, which has 16.9%UFR) that
239 has a single corresponding contig (ctg1519) with the same four genes (Supplementary Table 1).
240 We assigned ctg2522 to the proto-Y chromosome and ctg1519 to the proto-X (Figure 2B). All
241 four genes shared by both contigs are mapped to the third chromosome in the reference genome.
242 These four genes are present in the same order and same orientation in the reference genome (on
243 scaffold NW_004764689.1), the proto-Y contig, and proto-X contig.
244
245 One of the three genes on the *Mdmd*-carrying contig (ctg2382) and all four genes on the other
246 proto-Y contig (ctg2522) are nuclear-encoded mitochondrial genes (Supplementary Table 2).
247 Nuclear-encoded mitochondrial genes can evolve under sexually antagonistic selection (Rand et
248 al. 2001), possibly because of conflicts over mitochondrial functions in sperm and in other
249 tissues (Gemmell et al. 2004; Gallach et al. 2010). These inter-sexual conflicts could favor or
250 select against X-linked mitochondrial genes depending on the extent of co-transmission of the X

251 chromosome and mitochondria in females (Drown et al. 2012; Dean et al. 2014). Our results
252 suggest that mito-nuclear inter-sexual conflicts might also be important for X-Y divergence in
253 young sex chromosomes, where sexually antagonistic variants could be partitioned between
254 female-beneficial X-linked alleles and male-beneficial Y-linked alleles. Additional work is
255 required to test this hypothesis.
256
257 Two of the genes on the proto-Y contigs have differences in their protein-coding sequence from
258 their proto-X chromosome gametologs. One of the two (LOC101894698, encoding fast kinase
259 domain-containing protein 5, mitochondrial) contains three missense variable sites at which
260 genotypic (III^M/III) males are heterozygous and sex-reversed (III/III) male are homozygous
261 (Supplementary Table 3). The protein encoded by this gene contains an RNA-binding (RAP)
262 domain, but all three missense variants are not found in the domain. The other gene
263 (LOC101893231, which does not have a predicted mitochondrial function) has three missense
264 alleles at which genotypic males are heterozygous and sex-reversed male are homozygous
265 (Supplementary Table 3). One of three missense sites (at position 29,021 in the scaffold of the
266 reference genome) in this gene is found in a proline aminopeptidase P II domain. For all
267 missense alleles in both genes, we inferred the III allele as the one in common between
268 genotypic and sex-reversed males and the III^M allele as the one unique to genotypic males.
269 However, the inferred III^M alleles are not specific to the III^M chromosome because all III^M alleles
270 in LOC101894698 and one of the III^M alleles in LOC101893231 (at position 29,021) were found
271 in the reference genome (which comes from a genotypic female without a III^M chromosome).
272 Therefore, some of the III^M alleles in these genes are segregating as polymorphic variants on the
273 proto-X chromosome.
274
275 An alternative explanation of ctg2382 and ctg2522 is that they are paralogous sequences that
276 have been duplicated on the proto-Y chromosome, creating a second Y-linked copy of the genes
277 within the duplicated region. Intra-chromosomal duplications are a common feature of Y
278 chromosomes (Skaletsky et al. 2003; Hughes et al. 2010; Soh et al. 2014; Bachtrog et al. 2019;
279 Ellison and Bachtrog 2019). It is therefore possible that the sequences we classify as on the
280 proto-X are actually present on both the proto-X and proto-Y, while the contigs we classify as
281 proto-Y sequences are intra-chromosomal duplications on the proto-Y. However, even in this

282 scenario, the proto-Y sequences are still unique to the proto-Y. Therefore, our interpretations are
283 unlikely to be affected by whether the sequences we classify as proto-Y are true gametologs of
284 the proto-X or intra-chromosomal duplications on the proto-Y.
285
286 **Gene expression divergence between the proto-Y and proto-X chromosomes**
287
288 We next tested if differences in *cis*-regulatory sequences between the proto-Y and proto-X
289 chromosomes contribute to expression differentiation. We quantified differential expression
290 between the proto-X and proto-Y chromosome copies of the third chromosome genes by
291 measuring allele-specific expression (ASE) in normal (genotypic) males carrying a III^M proto-Y
292 chromosome and sex-reversed (III/III) males with no proto-Y chromosome. Comparing
293 genotypic and sex-reversed males allows us to control for the effect of sexually dimorphic gene
294 expression on the inference of divergence between the proto-Y (III^M) and proto-X (III)
295 chromosomes.
296
297 To quantify ASE of genes in genotypic (III^M/III) and sex-reversed (III/III) males, we used
298 existing RNA-seq data (Son et al. 2019) along with our new Oxford Nanopore long read
299 sequencing data. We used the IDP-ASE pipeline (Deonovic et al. 2017), which is more accurate
300 when a hybrid of short and long reads is provided as input. This is because the long reads are
301 used to construct haplotypes, which are used to estimate ASE from the RNA-seq data (see
302 Materials and Methods). This approach differs from our comparison of proto-X and proto-Y
303 contigs because the ASE analysis does not require separate assembly of proto-X and proto-Y
304 contigs. Instead, IDP-ASE uses the raw long read sequences from genotypic and sex reversed
305 males to phase haplotypes when inferring ASE. We measured ASE as the proportion of iterations
306 in a Markov chain Monte Carlo (MCMC) simulation in which the expression of a focal
307 haplotype is estimated as >0.5. This proportion gives a measure of ASE ranging from 0 (extreme
308 ASE in favor of one allele) to 1 (extreme ASE in favor of another allele), with 0.5 indicating
309 equal expression of both alleles. We are unable to determine if the focal haplotype refers to the
310 III^M or III allele across the entire third chromosome, but we do differentiate between these alleles
311 for a handful of genes where we are able to perform manual curation (see below).
312

313 We assigned each gene with sufficient expression data into one of five bins of ASE. The
314 proportions of iterations with focal haplotypes >0.5 were overrepresented at five values (0, 0.25,
315 0.5, 0.75, and 1) both in genotypic and sex-reversed males (Supplementary Figure 2 and 3).
316 These proportions may be overrepresented because we only sampled two genotypes for our ASE
317 analysis, which caused us to have a non-continuous distribution of proportions. We divided the
318 proportion of iterations with the focal haplotypes >0.5 into five bins, with each bin capturing one
319 of the five most common proportions (Supplementary Figure 2 and 3): 1) extreme ASE, with a
320 value between 0 and 0.125; 2) moderate ASE, with a value between 0.125 and 0.375; 3) non-
321 ASE, with a value between 0.375 and 0.625; 4) moderate ASE, with a value between 0.625 and
322 0.875; and 5) extreme ASE, with a value between 0.875 and 1. In the analysis below, we
323 considered a gene to have ASE if it falls into one of the two bins of extreme ASE; genes in the
324 non-ASE bin (proportion of focal haplotype >0.5 between 0.375 and 0.625) were classified as
325 having non-ASE. Genes with moderate ASE were excluded from most of our analyses in order to
326 be conservative about ASE assignment.
327
328 We first investigated ASE of genes we identified at the two loci where we have separate proto-X
329 and proto-Y contigs from our genome assembly (Supplementary Table 2). We had enough
330 sequence coverage and heterozygous sites to estimate ASE in 5/7 genes. Two mitochondrial
331 genes (LOC101894537 and LOC101894698) had extreme ASE in genotypic males and moderate
332 ASE in sex-reversed males. The elevated ASE in genotype males is suggestive of expression
333 divergence between the proto-Y and proto-X copies. Three genes (LOC101892763,
334 LOC101893231, and LOC101894024) exhibited extreme ASE in sex-reversed males and non-
335 ASE or moderate ASE in genotypic males. Elevated ASE in sex-reversed males is not expected
336 because they are genotypic females with two copies of the proto-X chromosome. This result
337 suggests there is not large-scale expression divergence between the proto-Y and proto-X
338 chromosomes. The limited expression divergence (Supplementary Table 2) and paucity of fixed
339 amino acid differences (Supplementary Table 3) between proto-Y and proto-X gametologs are
340 consistent with minimal differentiation between the house fly proto-Y and proto-X
341 chromosomes.
342

343 We next examined ASE across the entire third chromosome. If the III^M proto-Y chromosome is
344 differentiated in gene expression from its homologous III proto-X chromosome because of
345 differences in *cis*-regulatory alleles across the entire third chromosome, then we expect a higher
346 fraction of genes with ASE on the third chromosome in the genotypic (III^M/III) males than in the
347 sex-reversed (III/III) males. In contrast to that expectation, we did not find an excess of genes
348 with ASE in genotypic males compared to ASE genes in sex-reversed males on the third
349 chromosome relative to other chromosomes (Figure 3A and Supplementary Table 4; Fisher's
350 exact test, $P = 0.6996$). This result suggests that the III^M proto-Y chromosome is not broadly
351 differentiated in *cis*-regulatory alleles from the standard third (proto-X) chromosome. This
352 provides evidence that the early stages of Y chromosome evolution do not involve chromosome-
353 wide changes in gene regulation via an excess of *cis*-regulatory mutations.
354
355 We next identified individual genes with differences in ASE between genotypic (III^M/III) and
356 sex-reversed (III/III) males. There are 95 third chromosome genes with ASE in the genotypic
357 males that are non-ASE in the sex-reversed males (Supplementary Table 5). These genes could
358 have ASE in III^M males because of differences in *cis* regulatory sequences between the III^M and
359 standard third chromosome. To test whether the observed number of third chromosome genes
360 with ASE in genotypic males that are non-ASE in sex-reversed males is in excess of a null
361 expectation, we determined the number of third chromosome genes with ASE in sex-reversed
362 males that are non-ASE in genotypic males (i.e., the opposite of what we did above to find the
363 first set of 95 genes). There are 76 third chromosome genes with ASE in the sex-reversed males
364 that are non-ASE in genotypic males (Supplementary Table 5). We also identified 241 genes on
365 other chromosomes with ASE in genotypic males that are non-ASE in sex-reversed males, as
366 well as 281 genes on other chromosomes with ASE in sex-reversed males that are non-ASE in
367 genotypic males (Figure 3B and Supplementary Table 5). We do not expect any difference in
368 ASE between genotypic and sex reversed males for chromosomes other than the third.
369 Comparing genes with and without ASE on the third chromosome and the rest of the genome,
370 there is indeed an excess of third chromosome genes with ASE in genotypic males that are non-
371 ASE in sex-reversed males (Figure 3B and Supplementary Table 5; Fisher's exact test, $P =$
372 0.03467). These results suggest that, while the III^M proto-Y chromosome is not broadly
373 differentiated in *cis*-regulatory sequences from the standard third (proto-X) chromosome, there is

374 an excess of individual genes with *cis*-regulatory differences between the III^M proto-Y and its
375 homologous proto-X chromosome.
376
377 Male-specific selection on individual proto-Y genes could be responsible for ASE in genotypic
378 III^M males. In this scenario, male-specific selection would favor *cis*-regulatory variants that drive
379 up- or down-regulation of the III^M copy of a gene (Parsch and Ellegren 2013; Mank 2017). These
380 sex-specific selection pressures are expected to have the greatest effect for genes closest to the
381 male-determining *Mdmd* locus because those genes are most likely to be co-inherited with *Mdmd*
382 (Charlesworth et al. 2014). Unfortunately, we lack a chromosome-scale assembly of the house
383 fly genome (Scott et al. 2014; Meisel and Scott 2018), which prevents us from testing if genes
384 with ASE are clustered on the third chromosome in close proximity to the *Mdmd* locus.
385
386 It is also worth noting that read mapping bias can affect inference of ASE involving sex
387 chromosomes (Zimmer and Mank 2016; Wei and Bachtrog 2019). The house fly reference
388 genome was sequenced from female DNA (Scott et al. 2014), and it only has sequences derived
389 from the proto-X copy of the third chromosome. Proto-X derived reads should therefore
390 preferentially map to the reference genome (relative to proto-Y reads) if there is sequence
391 divergence between the proto-X and proto-Y chromosomes. Reduced mapping of proto-Y reads
392 should decrease our power to detect ASE in genotypic males. However, we find that the allele
393 we assign to the III^M proto-Y in the strain we have analyzed is frequently present in the female
394 reference genome (containing only proto-X sequences) generated from a different strain
395 (Supplementary Table 3; see Materials and Methods), suggesting that there may not be a
396 systemic mapping bias in our analysis. Regardless, we can conclude that our identification of an
397 excess of individual genes with ASE in genotypic males and non-ASE in sex reversed males
398 (Figure 3B) is not attributable to mapping bias because it is in the opposite direction expected to
399 be caused by mapping biases.
400
401 **Up-regulation of the Y-allele and male-biased expression of a testis gene**
402
403 We next tested if genes with ASE are differentially expressed between genotypic males, sex-
404 reversed males, and females. A relationship between ASE and sexually dimorphic expression

405 would suggest that up- or down-regulation of Y-linked alleles affects sexually dimorphic
406 phenotypes. We started by selecting genes that we had previously identified as differentially
407 expressed between genotypic (III^M/III) and sex-reversed (III/III) males (Son et al. 2019). These
408 two genotypes have nearly the same expression profiles, with a small number of differentially
409 expressed genes. We were specifically interested in genes on the third (proto-sex) chromosome
410 with "discordant sex-biased expression". Discordant sex-biased genes have male-biased
411 expression in genotypic males (i.e., upregulated relative to phenotypic females) and female-
412 biased expression in the sex-reversed males (downregulated relative to phenotypic females), or
413 vice versa. This pattern of expression is suggestive of *cis*-regulatory divergence between the
414 proto-Y and proto-X chromosomes—a hypothesis that was presented but not tested in our
415 previous study (Son et al. 2019). Here, we test that hypothesis by determining if any third
416 chromosome genes with discordant sex-biased expression have ASE consistent with *cis*-
417 regulatory divergence between the proto-Y (III^M) and proto-X (III) alleles.
418
419 We identified a single gene (LOC101899975) with discordant sex-biased gene expression out of
420 the 95 genes on the third chromosome with ASE in the genotypic (III^M/III) males that are non-
421 ASE in the sex-reversed (III/III) males. This gene is homologous to *dynein assembly factor 5,*
422 *axonemal* (human gene *DNAAF5* and *Drosophila melanogaster* gene *HEATR2*). The gene, which
423 we refer to as *M. domestica HEATR2* (*Md-HEATR2*), is expected to encode a protein that
424 functions in flagellated sperm motility (Diggle et al. 2014), and it has strong testis-biased
425 expression in *D. melanogaster* (Chintapalli et al. 2007). *Md-HEATR2* has male-biased expression
426 in the abdomens of genotypic males and female-biased expression in the abdomens of sex-
427 reversed males (Son et al. 2019), suggesting that expression differences between the III^M proto-Y
428 and the standard third (proto-X) chromosome cause the male-biased expression of the gene in the
429 genotypic males.
430
431 With haplotypes estimated by IDP-ASE, we identified three diagnostic variant sites for ASE
432 within *Md-HEATR2* (Figure 4A), which are all synonymous SNPs. The genotypic (III^M/III)
433 males are heterozygous and the sex-reversed (III/III) males are homozygous at all diagnostic
434 sites. We inferred the allele on the standard third chromosome as the one in common between
435 genotypic and sex-reversed males, and the III^M allele as the one unique to genotypic males at

436 each diagnostic variant site. Curiously, all three III^M alleles are found in the reference genome,
437 suggesting that these synonymous variants are not fixed differences between the proto-Y and
438 proto-X. *Md-HEATR2* is expressed higher in III^M genotypic males than in sex-reversed males
439 (Figure 4A). In the III^M genotypic males, the III^M (Y-linked) alleles are expressed higher than the
440 X-linked alleles, indicating that the Y-linked alleles are associated with the up-regulation of the
441 gene in III^M genotypic males relative to sex-reversed males (Figure 4A). The copy of *Md-*
442 *HEATR2* on the III^M proto-Y chromosome is therefore up-regulated relative to the proto-X copy,
443 consistent with higher expression of *Md-HEATR2* in genotypic males.
444
445 Using our Nanopore sequencing reads mapped to the reference genome, we examined 1,273 base
446 pairs upstream of *Md-HEATR2* to identify diagnostic sites that could be responsible for
447 regulating the expression differences between the proto-X and proto-Y alleles. We chose that
448 distance because it includes the first variable site we could identify on the scaffold containing
449 *Md-HEATR2* in our Nanopore data (i.e., including a larger region would not provide any
450 additional information). We found twelve variable sites with different alleles (SNPs and small
451 indels) between genotypic (III^M/III) and sex-reversed (III/III) males (Figure 4B). We next
452 examined whether these sites are located within a potential transcription factor (TF) binding
453 region. We found five TF binding regions predicted upstream of *Md-HEATR2* using the
454 'Tfsitescan' tool in the 'object-oriented Transcription Factors Database' (Ghosh 2000). However,
455 none of the twelve variable sites are found within any predicted TF binding regions (Figure 4C).
456 It is possible that the *cis*-regulatory sequences responsible for differential expression are located
457 outside of the region we were able to investigate, which would be consistent with our failure to
458 identify fixed differences in the exons between the proto-Y and proto-X copies of *Md-HEATR2*.
459 A long distance would reduce the genetic linkage between the *cis*-regulatory region and
460 transcribed sequence, allowing for the X-Y differentiation of the regulatory region without
461 differentiation of the transcribed gene. Further work is needed to determine how the differential
462 expression of the proto-X and proto-Y copies of *Md-HEATR2* is regulated.
463
464 We hypothesize that the upregulation of the proto-Y copy of *Md-HEATR2* is the result of
465 selection for higher expression in males. We think that the *cis*-regulatory region is under
466 selection, rather than the protein-coding sequence, because we fail to find any protein coding

467 differences between the proto-X and proto-Y copies, and the synonymous differences we
468 observe are not fixed differences. HEATR2 is involved in dynein arm assembly (Diggle et al.
469 2014), and axonemal dynein is essential for flagellated sperm motility (Kurek et al. 1998;
470 Carvalho et al. 2000). Therefore, it may be beneficial to male fitness to have higher expression of
471 *Md-HEATR2* in testis. However, HEATR2 also functions in mechanosensory neurons in
472 *Drosophila* (Diggle et al. 2014). There may be conflict over the *cis*-regulatory sequences that
473 promote expression in testis and neurons, which may prevent up-regulation of the proto-X copy
474 of *Md-HEATR2* in testis. These opposing (i.e., sexually antagonistic) selection pressures on *Md-*
475 *HEATR2* expression would be resolved in a Y-linked copy that is only under selection in males
476 (Rice 1996b). Alternatively, the expression differences between sex-reversed and genotypic
477 males could be caused by differences in tissue scaling between genotypes (Montgomery and
478 Mank 2016). For example, if genotypic males have larger testes and if the proto-Y allele is
479 preferentially expressed in testis, then it will appear as if genotypic males have higher expression
480 of *Md-HEATR2*. Even in this model, there is ASE associated with testis-specific expression,
481 which is consistent with male-specific selection favoring the proto-Y allele.
482
483 The up-regulation or testis-biased expression of the proto-Y copy of *Md-HEATR2* may not
484 completely resolve sexually antagonistic selection pressures on *Md-HEATR2* expression because
485 of three notable features of house fly genetics. First, Y-linked alleles are only under male-
486 specific selection when they are in complete genetic linkage with a male-determining gene
487 (Charlesworth et al. 2014). Complete linkage between loci can be caused by recombination
488 suppressors (e.g., chromosomal inversions) that prevent genetic exchange between the X and Y
489 chromosomes (Rice 1987; Charlesworth 2017). While there is no direct evidence for inversions
490 or other suppressors of recombination on the house fly third chromosome, crossing over in male
491 meiosis is rare in most flies (Gethmann 1988). A general lack of recombination in males would
492 prevent X-Y exchange. However, male recombination has been documented in house fly
493 (Feldmeyer et al. 2010), suggesting that X-Y exchange is possible. Second, female house flies
494 can carry proto-Y chromosomes if they also carry the female-determining *Md-tra*^{D} allele (Hamm
495 et al. 2015). Female-transmission of the proto-Y provides another avenue for X-Y exchange in
496 the absence of an inversion or other suppressor of recombination on the proto-Y or proto-X. We
497 observe alleles in *Md-HEATR2* and other third chromosome genes that are found on both the

498 proto-Y and proto-X (Supplementary Table 3), which is consistent with either X-Y
499 recombination or ancestral polymorphisms that predate the formation of the proto-sex
500 chromosome. Third, when females carry a proto-Y chromosome, there will be selection against
501 male-beneficial, female-detrimental sexually antagonistic alleles, which will further prevent the
502 resolution of sexual conflict.
503
504 Selection on proto-Y alleles in females, along with X-Y recombination that moves proto-X
505 alleles onto the proto-Y, will reduce the efficacy of male-specific selection pressures on sexually
506 antagonistic alleles of *Md-HEATR2*. The extent of these effects will depend on the frequency of
507 proto-Y chromosomes found in females. First, the extent to which selection in females acts
508 against male-beneficial sexually antagonistic alleles depends on the frequency with which the
509 proto-Y chromosomes are found in females (Rice 1984). Second, if recombination rate is
510 sexually dimorphic, the rate of population level recombination between the male-determiner and
511 *Md-HEATR2* will depend on how often the proto-Y chromosomes are found in females (Sardell
512 et al. 2018). Notably, the frequencies of proto-Y chromosomes and the female-determining *Md-*
513 tra^D allele vary across populations (Hamm et al. 2015; Meisel et al. 2016). Therefore, the
514 frequency with which females carry proto-Y chromosomes will vary across populations,
515 suggesting that selection against male-beneficial *cis*-regulatory alleles of *Md-HEATR2* on the
516 proto-Y may be population-specific.
517
518 The expression divergence of the proto-X and proto-Y copies of *Md-HEATR2* could constitute
519 an early stage of X-Y differentiation before chromosome-wide X-Y differentiation occurs
520 (Bachtrog 2013). Young Y chromosomes have very similar gene content as their ancestral
521 autosomes. In contrast, old Y chromosomes are more likely to have only retained genes with
522 male-specific functions or recruited genes associated with testis expression from other autosomes
523 (Koerich et al. 2008; Kaiser et al. 2011; Mahajan and Bachtrog 2017). Our results suggest that
524 changes in the expression of individual Y-linked genes that were retained from the ancestral
525 autosome could have important phenotypic effects during early Y chromosome evolution,
526 consistent with gene-by-gene divergence on the young Y chromosomes (Wei and Bachtrog
527 2019).
528

529 **Conclusions**

530

531 We investigated gene sequence and expression differences between the house fly III^M proto-Y
532 chromosome and its homologous proto-X to determine how a very young proto-Y/proto-X pair
533 diverge shortly after the proto-Y was formed. To those ends, we used genotypic (III^M/III) and
534 sex-reversed (III/III) males because they are phenotypically almost the same and only differ in
535 whether they carry a proto-Y chromosome (Hediger et al. 2010; Son et al. 2019). We observe
536 elevated heterozygosity on the proto-sex chromosome in genotypic males (Figure 1), which
537 could be indicative of chromosome-wide sequence differentiation between the proto-Y and
538 proto-X. Alternatively, elevated heterozygosity in genotypic males could merely be a result of
539 being heterozygous for the third chromosome. Consistent with this alternative hypothesis, when
540 we previously substituted a third chromosome without *Mdmd* onto a common genetic
541 background, the effect on gene expression was comparable to substituting a III^M chromosome on
542 the same background (Son et al. 2019).

543

544 Our subsequent analyses suggest that the house fly III^M proto-Y chromosome is differentiated in
545 sequence and expression from its homologous proto-X chromosome at individual genes, but not
546 chromosome-wide. This is consistent with previous work that found evidence for gene-by-gene
547 divergence in a young neo-Y chromosome (Wei and Bachtrog 2019). For example, we only
548 identified two genomic loci (containing a total of 7 genes) that are sufficiently differentiated to
549 assemble into separate proto-X and proto-Y contigs (Figure 2). Notably, 5/7 of those genes have
550 mitochondrial functions, suggesting that mitochondrial function in sperm might be an important
551 target of male-specific selection during the early evolution of a Y chromosome. We did not
552 identify any fixed differences in the protein-coding sequences between the proto-X and proto-Y
553 gametologs of the five mitochondrial genes (Supplementary Table 3), or in the sequence of *Md-*
554 *HEATR2* (Figure 4), providing further evidence for minimal differentiation between the proto-X
555 and proto-Y. This also suggests that sex-specific selection may be operating on *cis* regulatory
556 sequences of these genes. Consistent with this hypothesis, two of the mitochondrial genes are
557 differentially expressed between the proto-Y and proto-X (Supplementary Table 2). Despite
558 these intriguing examples, there is only a moderate excess of genes with evidence for differential
559 expression between the proto-Y and proto-X (Figure 3). The number of genes with ASE on the

560 third chromosome only in genotypic males, and not in sex-reversed males, is small; we identify
561 less than 100 genes that meet these criteria, which is less than 5% of the 2,524 genes on the third
562 chromosome (Meisel and Scott 2018).
563
564 We identified one gene on the third chromosome (*Md-HEATR2*) with ASE in genotypic males
565 that is non-ASE in sex-reversed males and has discordant sex-biased expression between
566 genotypic and sex-reversed males (Figure 4). We hypothesize that expression divergence of *Md-*
567 *HEATR2* could be an example of very early X-Y differentiation of individual genes that results
568 from sex-specific selection. Notably, *Md-HEATR2* is expected to have important functions in
569 spermatogenesis or sperm motility (Diggle et al. 2014; Fuller 1993), similar to the five
570 mitochondrial genes contained within genomic regions that are divergent between the proto-Y
571 and proto-X. Male fertility is an important target of selection during the evolution of old Y
572 chromosomes (Carvalho et al. 2009; Hughes et al. 2010), and our results suggest that selection
573 on male fertility is also an important driver of X-Y divergence during the early evolution of sex
574 chromosomes. Our results suggest that these selection pressures during earliest stages of Y
575 chromosome evolution drive gene-by-gene, rather than chromosome-scale, changes in gene
576 expression.
577
578
579 **Materials and Methods**
580
581 **<u>Fly strain</u>**
582
583 We analyzed RNA-seq data and performed Oxford Nanopore sequencing on a house fly strain
584 that allows for identification of genotypic (III^M/III) males and sex-reversed (III/III) males
585 (Hediger et al. 2010). This is because the standard third chromosome (III) in this strain has the
586 recessive mutations *pointed wing* (*pw*) and *brown body* (*bwb*). Sex-reversed males (and normal
587 females) have both mutant phenotypes, whereas genotypic males are wild-type for both
588 phenotypes because the III^M chromosome has the dominant wild-type alleles. The RNA-seq data
589 that we analyzed (available at NCBI GEO accession GSE126689) comes from a previous study
590 that used double-stranded RNA (dsRNA) targeting *Md-tra* to create sex-reversed phenotypic

591 males that have a female genotype without a proto-Y chromosome (Son et al. 2019). This is
592 because active *Md-tra* drives female development and inactive *Md-tra* triggers male
593 development (Hediger et al. 2010). We compared gene expression in the abdomens of sex-
594 reversed males, genotypic males that received a sham treatment of dsRNA targeting GFP, and
595 genotypic females that are phenotypically female. We analyzed data from three replicates of each
596 genotype and sex, with each replicate consisting of a single fly abdomen (Son et al. 2019). We
597 used genotypic males and sex-reversed males from the same strain that were subjected to the
598 same treatment for genome sequencing using the Oxford Nanopore long read technology.
599
600 **Oxford Nanopore sequencing**
601
602 We performed Oxford Nanopore sequencing of one genotypic (III^M/III) male and one sex-
603 reversed (III/III) male created from the same strain and using the same *Md-tra* dsRNA treatment
604 as a previous RNA-seq study (Son et al. 2019). DNA was isolated with a phenol/chloroform
605 protocol. A single genotypic male and a single sex-reversed male with detached wings were each
606 transferred to a 1.5 mL Eppendorf tube with 0.5 mL homogenization buffer (4.1 g sucrose, 15
607 mL 1M Tris-HCl pH 8.0, 0.5M EDTA, 100 mL dH_2O) and then homogenized using pestles set
608 into a tissue grinder homogenizer. To each tube we added 40 uL of 10% SDS and 2.5 uL of
609 10mg/mL Proteinase K, and then we incubated the tube at 65°C for 30 min. We next added 2 uL
610 of 4 mg/mL RNase to each tube and incubated at 37°C for 15 min. We added 48 uL of 5M KAc
611 to each tube and placed on ice for 30 min. Then we centrifugated the tubes at 14000 rpm for 10
612 min at 4°C, and the supernatant was transferred into a new tube using a wide-bore pipette tip (all
613 subsequent steps of transfer and mixing during DNA extraction were also done with wide-bore
614 pipette tips to prevent DNA shearing). We added 250 uL phenol and 250 uL chloroform to the
615 extracted supernatant in the new tube, mixed briefly, spun at 14000 rpm for 15 min at 4°C, and
616 then transferred the supernatant into a new tube. We next added 500 uL chloroform to the
617 supernatant in the new tube, mixed well, and spun a 14000 rpm for 5 min at 4°C. We then
618 transferred the supernatant into a new tube. We added 40 uL of 3M NaAc and 800 uL of 95%
619 ethanol to the supernatant in the new tube, mixed briefly, spun at 14000 rpm for 15 min at 4°C,
620 and then carefully poured off all supernatant. We next added 800 uL of 70% ethanol to the

621 remaining pellet, mixed briefly to wash the pellet, spun at 14000 rpm for 15 min at 4°C, removed
622 the supernatant, and then resuspended the pellet in 30 uL of nuclease-free water.
623
624 Oxford Nanopore Sequencing libraries were prepared with the 1D genomic DNA Ligation kit
625 (SQK-LSK109, Oxford Nanopore), following the manufacturer's protocol. DNA from the
626 genotypic male and sex-reversed male (see above) was used to create a separate sequencing
627 library for each genotype. Following the manufacturer's protocol, 15 uL of each library, along
628 with sequencing buffer and loading beads (totaling 75 uL), were separately loaded onto two
629 different R9.4 flow cells (i.e., the libraries from the genotypic and sex-reversed males were run
630 on separate flow cells) until no pores were available on a MinION sequencer (Oxford Nanopore).
631 We performed two different base calling pipelines, using the Guppy pipeline software version
632 3.1.5 (Oxford Nanopore). First, for the IDP-ASE analysis, we used parameter options with "--
633 calib_detect --qscore_filtering --min_qscore 10". Second, for the genome assembly we used the
634 default parameters. We used different parameters for IDP-ASE because base quality affects
635 accurate haplotyping used to estimate ASE (see below), and we therefore used a higher threshold
636 for the base quality. For the IDP-ASE analysis, the base called reads were aligned to the house
637 fly genome assembly v2.0.2 (Scott et al. 2014) using Minimap2 version 2.17 with the "-ax map-
638 ont" parameter (Li 2018).
639
640 **Genome assembly, transcript alignment, and sequence divergence**
641
642 We used wtdbg2 to assemble our base called Oxford Nanopore reads using the default
643 parameters (Ruan and Li 2020). To find genic regions in our genome assembly, we aligned
644 house fly transcripts (from Annotation Release 102) as a query against our genome assembly
645 contigs using BLAT with a minimum mapping score of 50 (Kent 2002). We considered BLAT
646 alignments to be gene copies only if the matched sequence length in the BLAT alignments
647 covers at least half of the original (annotated) transcript length. We selected contigs in our
648 assembly that contain genes that are assigned to the third chromosome in the reference genome.
649
650 We used two approaches to differentiate the proto-Y and proto-X contigs. In both approaches we
651 specifically focused on pairs of contigs that contain the same genes because they are indicative

652 of X-Y divergence (i.e., one contig likely contains a proto-Y sequence, and the other contains a
653 proto-X sequence). First, we tested if one contig contains a copy of *Mdmd*, which allows us to
654 designate that contig as the proto-Y copy (and the other is proto-X). *Mdmd* is not present in the
655 annotated house fly genome because the genome was sequenced from female DNA (Scott et al.
656 2014). To find *Mdmd* copies in our genome assembly, we used the sequence of the *Mdmd*
657 transcript (Sharma et al. 2017) as a query to find contigs containing *Mdmd* with BLAST using an
658 e-value cutoff of 1e-3 (Altschul et al. 1990). Second, we used a *k*-mer comparison approach to
659 differentiate the proto-X and proto-Y contigs (Carvalho and Clark 2013). In this approach, we
660 used a *k*-mer size of 15 to measure the %UFR for every contig in our assembly, with higher
661 values indicating an increased likelihood that a contig is Y-linked. The female reads were from
662 the genome project (BioProject accession PRJNA176013; Scott et al. 2014), and we included the
663 III^M male Oxford Nanopore sequencing reads we generated for validation of the bit-array. We
664 followed the options suggested to identify Y sequences in *Drosophila* genomes (Carvalho and
665 Clark 2013), as described previously (Meisel et al. 2017). We used the R package Gviz (Hahne
666 and Ivanek 2016) to visualize genes that are present in III^M (proto-Y) contigs and III (proto-X)
667 contigs.
668
669 **Variant calling**
670
671 We used available RNA-seq data (Son et al. 2019) to identify genetic variants (SNPs and small
672 insertions/deletions) that differentiate the III^M proto-Y chromosome from the standard third
673 (proto-X) chromosome, and then we tested if III^M males have elevated heterozygosity on the
674 third chromosome as compared to sex-reversed males (Meisel et al. 2017). We used the Genome
675 Analysis Toolkit (GATK) pipeline for calling variants in the RNA-seq data from the *Md-tra*
676 RNAi experiment in (Son et al. 2019), following the best practices for SNP and
677 insertion/deletion (indel) calling on RNA-seq data (McKenna et al. 2010; Meisel et al. 2017). We
678 used STAR (Dobin et al. 2013) to align reads from three genotypic (III^M/III) male libraries and
679 three sex-reversed (III/III) male libraries to the (female) reference assembly v2.0.2 (Scott et al.
680 2014). The reference genome was sequenced from the aabys strain, which differs from the one
681 we used in our experiment and has males that carry the Y^M proto-Y chromosome. The aligned
682 reads were used to generate a new reference genome index from the detected splice junctions in

683 the first alignment run, and then a second alignment was performed with the new reference. We
684 next marked duplicate reads from the same RNA molecule and used the GATK tool
685 'SplitNCigarReads' to reassign mapping qualities to 60 with the 'ReassignOneMappingQuality'
686 read filter for alignments with a mapping quality of 255. Indels were detected and realigned with
687 'RealignerTargetCreator' and 'IndelRealigner'. The realigned reads were used for base
688 recalibration with 'BaseRecalibrator' and 'PrintReads'. The base recalibration was performed in
689 three sequential iterations in which recalibrated and filtered reads were used to train the next
690 round of base recalibration, at which point there were no beneficial effects of additional base
691 recalibration as verified by 'AnalyzeCovariates'. We next used the recalibrated reads from all
692 three replicates of genotypic and sex-reversed males to call variants using 'HaplotypeCaller'
693 with emission and calling confidence thresholds of 20. We applied 'genotypeGVCFs' to the
694 variant calls from the two types of males for joint genotyping, and then we filtered the variants
695 using 'VariantFiltration' with a cluster window size of 35 bp, cluster size of 3 SNPs, FS > 20,
696 and QD < 2. The QD filter simultaneously considers read depth and variant quality so that a
697 separate read depth filter is not needed during variant filtration. Because this filtration is applied
698 during the joint genotypic step, all variants are genotyped using the combined information across
699 both types of males, eliminating the need to separately cross-reference read mapping information
700 across genotypes. The final variant calls were used to identify heterozygous variants within
701 genes using the coordinates from the genome sequencing project, annotation release 102 (Scott et
702 al. 2014). If a locus has different variants (e.g. a reference allele and an alternative allele), we
703 considered the locus heterozygous. We measured relative heterozygosity within each gene in
704 genotypic (III^M/III) and sex-reversed (III/III) males as the number of heterozygous variants in
705 genotypic males for a given gene (h_G) divided by the total number heterozygous variants in both
706 genotypic and sex-reversed males (h_{SR}), times one hundred: $100h_G/(h_G + h_{SR})$. To annotate each
707 variant within coding sequences, we used SnpEff with filter options "-onlyProtein", "-no-
708 intergenic", "-no-downstream", "-no-upstream", "-no-intron", and "-no-utr" (Cingolani et al.
709 2012).
710
711 For the variant calling from Nanopore long reads, the base called reads were indexed using fast5
712 files with the 'index' module of Nanopolish version 0.11.1 (Quick et al. 2016), and they were
713 aligned with Minimap2 version 2.17 (Li 2018) to house fly genome assembly v2.0.2 (Scott et al.

714 2014). The aligned and raw reads were used to call variants using the "variants" module of
715 Nanopolish version 0.11.1 with the "--ploidy 2" parameter (Quick et al. 2016). We used a python
716 script 'nanopolish_makerange.py' provided in the package to split the genome into 50 kb
717 segments because it was recommended to use the script for large datasets with genome size more
718 than 50 kb.
719
720 **Allele-specific expression**
721
722 ASE is the unequal expression of the maternal and paternal alleles in a diploid. We investigated
723 if there is elevated ASE on the third chromosome in males carrying one III^M proto-Y and one
724 proto-X chromosome compared to sex-reversed males with two proto-X chromosomes. To do
725 this, we implemented the IDP-ASE tool at the gene level with house fly genome annotation
726 release 102 (Scott et al. 2014), following the developers' recommended analysis steps (Deonovic
727 et al. 2017). We first prepared information on the number and locations of variants within each
728 gene (SNPs and small indels). The IDP-ASE software was supplied with raw and aligned reads
729 created by RNA-seq (Son et al. 2019) and Nanopore sequencing, as well as variant calls in RNA-
730 seq reads created only by GATK. We only used the Nanopore reads for phasing haplotypes in
731 the IDP-ASE run, and not for variant calling, because there was less than 10× coverage across
732 the house fly genome (i.e., too low for reliable variant calling).
733
734 The prepared data from each gene was next run in an MCMC sampling simulation to estimate
735 the haplotype within each gene with a Metropolis-Hastings sampler (Bansal et al. 2008). The
736 software estimates the proportion of each estimated haplotype that contributes to the total
737 expression of the gene (ρ) from each iteration using slice sampling (Neal et al. 2003). A value of
738 $\rho=0.5$ indicates equal expression between two alleles, whereas $\rho<0.5$ or $\rho>0.5$ indicates ASE.
739 The MCMC sampling was run with a 1000 iteration burn-in followed by at least 500 iterations
740 where data were recorded. The actual number of iterations was automatically adjusted by the
741 software during the simulation to produce the best simulation output for quantifying ASE within
742 a gene. The IDP-ASE simulation generated a distribution of ρ for each gene across all post-burn-
743 in iterations, and then it calculated the proportion of iterations with $\rho > 0.5$. This proportion was
744 used to estimate the extent of ASE for each gene. For example, if all iterations for a gene have ρ

745 > 0.5, then the proportion is 1 and the gene has strong evidence for ASE of one allele. Similarly,
746 if all iterations for a gene have $\rho < 0.5$, then the proportion is 0 and the gene has strong evidence
747 for ASE of the other allele. In contrast, if half of the iterations have $\rho > 0.5$ and the other half
748 have $\rho < 0.5$, then the proportion is 0.5 and there is not any evidence for ASE. In our subsequent
749 analysis, we only included genes with at least 10 mapped reads combined across three RNA-seq
750 libraries from the genotype under consideration.
751
752 We used the output of IDP-ASE to compare expression of the III^M (proto-Y) and III (proto-X)
753 alleles in genotypic males. IDP-ASE only quantifies ASE within bi-allelic loci, so we only
754 included genes with heterozygous sites within transcripts in genotypic (III^M/III) or sex-reversed
755 (III/III) males. In addition, we removed heterozygous variants with the same genotype in
756 genotypic and sex-reversed males because they do not allow us discriminate between the proto-Y
757 and proto-X alleles. Removing these variants may have also sped up the simulation times, but
758 this was not rigorously investigated. To discriminate between the III^M and III alleles, we used
759 haplotypes estimated during IDP-ASE runs and genotypes inferred from GATK for genotypic
760 (III^M/III) and sex-reversed (III/III) males. For example, using genotypes called using GATK
761 from the RNA-seq data, we first identified sites with heterozygous alleles in genotypic males and
762 homozygous alleles in sex-reversed males. Next, we inferred the allele in common between
763 genotypic and sex-reversed as the III allele, and the other allele that is unique to genotypic males
764 as the III^M allele. Lastly, we matched those sites to the haplotypes estimated by IDP-ASE to
765 quantify ASE within each genotype.
766
767
768 **Acknowledgements and funding information**
769 We thank Leo W. Beukeboom and Dainel Bopp for the collaboration on the RNAi knockdown
770 and valuable discussions, and Fei Yuan for help developing a python script used to calculate the
771 percentage of heterozygous variants. This work was supported by a Grant-in-Aid of Research
772 from the National Academy of Sciences, administered by Sigma Xi, The Scientific Research
773 Society (grant number G2018100198487895 to J.H.S. and R.P.M.) and by the National Science
774 Foundation (grant numbers OISE-1444220 and DEB-1845686 to R.P.M.). Analysis of RNA-seq
775 and Oxford Nanopore sequencing data were performed on the Maxwell and Sabine clusters in

776 the Research Computing Data Core at the University of Houston. The Oxford Nanopore long
777 reads data used in the study are available from the National Center for Biotechnology
778 Information Sequence Read Archive under BioProject accession PRJNA620357 (BioSample
779 accessions SAMN14518459 for sex-reversed males and SAMN14518460 for genotypic males).
780 The III^M male genome assembly is available from the National Center for Biotechnology
781 Information Whole Genome Shotgun accession JACCFA000000000 under BioProject accession
782 PRJNA620357.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 14(2):113–124.

Bachtrog D, Mahajan S, Bracewell R. 2019. Massive gene amplification on a recently formed Drosophila Y chromosome. Nat Ecol Evol. 3(11):1587–1597.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? PLoS Biol. 12(7):e1001899.

Bansal V, Halpern AL, Axelrod N, Bafna V. 2008. An MCMC algorithm for haplotype assembly from whole-genome sequence data. Genome Res. 18(8):1336–1346.

Bergero R, Qiu S, Forrest A, Borthwick H, Charlesworth D. 2013. Expansion of the pseudo-autosomal region and ongoing recombination suppression in the Silene latifolia sex chromosomes. Genetics 194(3):673–686.

Beukeboom LW, Perrin N. 2014. The evolution of sex determination. Oxford University Press, USA.

Boyes JW, Corey MJ, Paterson HE. 1964. Somatic chromosomes of higher diptera: IX. Karyotypes of some muscid species. Can J Zool. 42(6):1025–1036.

Bull JJ. 1983. Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Inc.

Carvalho AB, Clark AG. 2013. Efficient identification of y chromosome sequences in the human and drosophila genomes. Genome Res. 23(11):1894–1907.

Carvalho AB, Koerich LB, Clark AG. 2009. Origin and evolution of Y chromosomes: Drosophila tales. Trends Genet. 25(6):270–277.

Carvalho AB, Lazzaro BP, Clark AG. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. Proc Natl Acad Sci. 97(24):13239–13244.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. Philos Trans R Soc London Ser B Biol Sci. 355(1403):1563–1572.

Charlesworth B, Jordan CY, Charlesworth D. 2014. The evolutionary dynamics of sexually antagonistic mutations in pseudoautosomal regions of sex chromosomes. Evolution (NY) 68(5):1339–1350.

Charlesworth D. 2017. Evolution of recombination rates between sex chromosomes. Philos Trans R Soc B Biol Sci. 372(1736):20160456.

Charlesworth D. 2018. Does sexual dimorphism in plants promote sex chromosome evolution? Environ Exp Bot. 146:5–12.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95(2):118.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet. 39(6):715–720.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6(2):80–92.

Darolti I, Wright AE, Sandkam BA, Morris J, Bloch NI, Farré M, Fuller RC, Bourne GR, Larkin DM, Breden F, et al. 2019. Extreme heterogeneity in sex chromosome differentiation and dosage

829 compensation in livebearers. Proc Natl Acad Sci U S A. 116(38):19031–19036.
830 Dean R, Zimmer F, Mank JE. 2014. The potential role of sexual conflict and sexual selection in
831 shaping the genomic distribution of mito-nuclear genes. Genome Biol Evol. 6(5):1096–1104.
832 Deonovic B, Wang Y, Weirather J, Wang X-J, Au KF. 2017. IDP-ASE: Haplotyping and
833 quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing.
834 Nucleic Acids Res. 45(5):e32--e32.
835 Diggle CP, Moore DJ, Mali G, zur Lage P, Ait-Lounis A, Schmidts M, Shoemark A, Garcia
836 Munoz A, Halachev MR, Gautier P, et al. 2014. HEATR2 Plays a Conserved Role in Assembly
837 of the Ciliary Motile Apparatus. PLoS Genet. 10(9):e1004577.
838 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
839 TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21.
840 Drown DM, Preuss KM, Wade MJ. 2012. Evidence of a paucity of genes that interact with the
841 mitochondrion on the X in mammals. Genome Biol Evol. 4(8):875–880.
842 Duronio RJ, Marzluff WF. 2017. Coordinating cell cycle-regulated histone gene expression
843 through assembly and function of the Histone Locus Body. RNA Biol. 14(6):726–738.
844 Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y
845 chromosomes. PLoS Genet. 15(7):e1008251.
846 Feldmeyer B, Pen I, Beukeboom LW. 2010. A microsatellite marker linkage map of the
847 housefly, Musca domestica: Evidence for male recombination. Insect Mol Biol. 19(4):575–581.
848 Finnegan DJ. 2012. Retrotransposons. Curr Biol. 22(11):R432–R437.
849 Fuller MT. 1993. Spermatogenesis. In: Bate M, Martinez-Arias A, editors. The Development of
850 Drosophila. Cold Spring Harbor, (NY): Cold Spring Harbor Laboratory Press. p. 71–147
851 Gallach M, Chandrasekaran C, Betran E. 2010. Analyses of nuclearly encoded mitochondrial
852 genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic
853 conflict in Drosophila. Genome Biol Evol. 2:835–850.
854 Garcia-Moreno J, Mindell DP. 2000. Rooting a phylogeny with homologous genes on opposite
855 sex chromosomes (gametologs): A case study using avian CHD. Mol Biol Evol. 17(12):1826–
856 1832.
857 Gemmell NJ, Metcalf VJ, Allendorf FW. 2004. Mother's curse: The effect of mtDNA on
858 individual fitness and population viability. Trends Ecol Evol. 19(5):238–244.
859 Gethmann RC. 1988. Crossing over in males of higher diptera (brachycera). J Hered. 79(5):344–
860 350.
861 Ghosh D. 2000. Object-oriented transcription factors database (ooTFD). Nucleic Acids Res.
862 28(1):308–310.
863 Gu L, Walters JR. 2017. Evolution of sex chromosome dosage compensation in animals: A
864 beautiful theory, undermined by facts and bedeviled by details. Genome Biol Evol. 9(9):2461–
865 2476.
866 Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and Bioconductor. In: Mathé E,
867 Davis S, editors. Statistical Genomics: Methods and Protocols. New York (NY): Springer. p.
868 335–351.
869 Hamm RL, Meisel RP, Scott JG. 2015. The evolving puzzle of autosomal versus Y-linked male
870 determination in Musca domestica. G3 (Bethesda) 5(3):371–384.
871 Hediger M, Henggeler C, Meier N, Perez R, Saccone G, Bopp D. 2010. Molecular
872 characterization of the key switch F provides a basis for understanding the rapid divergence of
873 the sex-determining pathway in the housefly. Genetics 184(1):155–170.
874 Hediger M, Minet AD, Niessen M, Schmidt R, Hilfiker-Kleiner D, Çakir Ş, Nöthiger R,

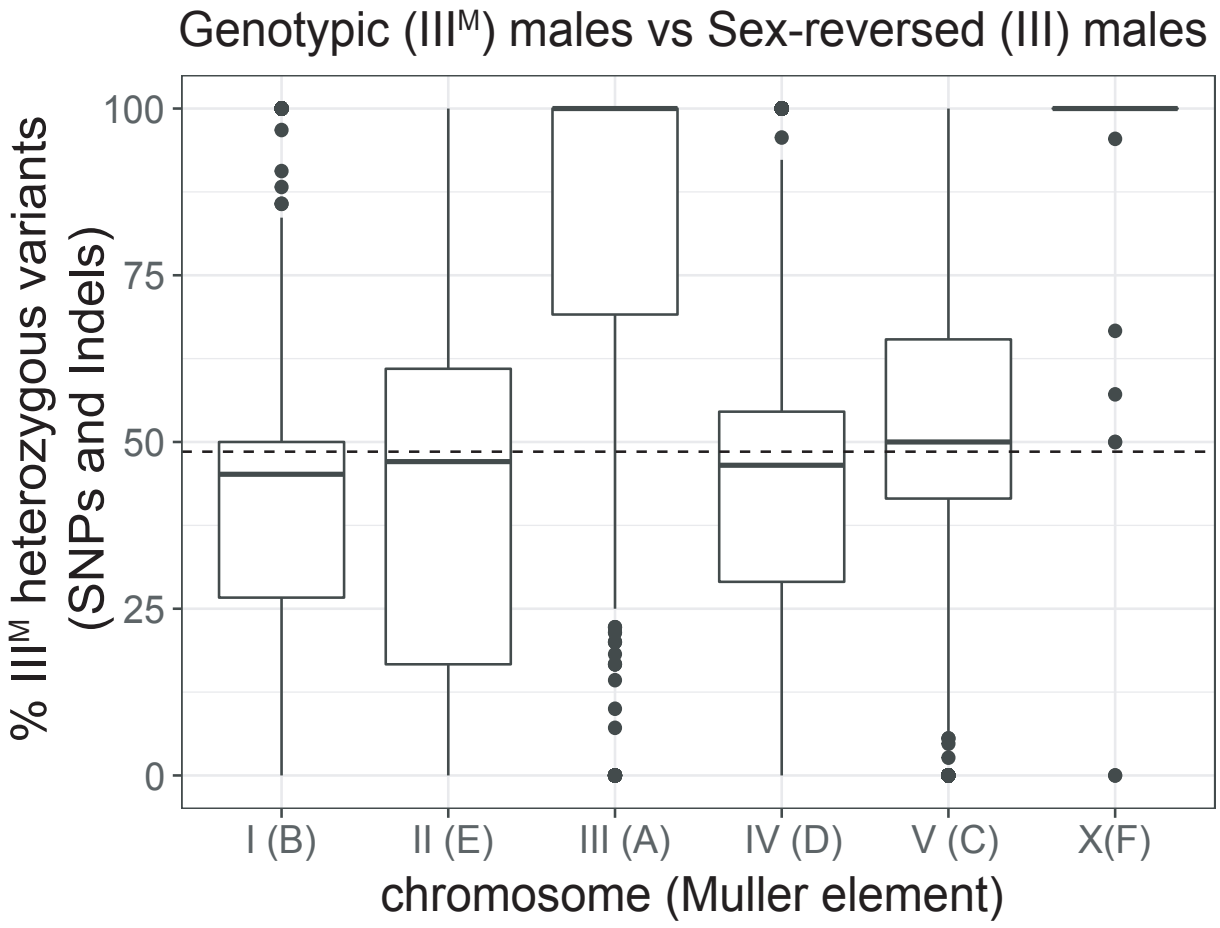
875 Dübendorfer A. 1998. The male-determining activity on the Y chromosome of the housefly
876 (*Musca domestica* L.) consists of separable elements. Genetics 150(2):651–661.
877 Hughes JF, Skaletsky H, Pyntikova T, Graves TA, Van Daalen SKM, Minx PJ, Fulton RS,
878 McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human y chromosomes are
879 remarkably divergent in structure and gene content. Nature 463(7280):536–539.
880 Kaiser VB, Zhou Q, Bachtrog D. 2011. Nonrandom gene loss from the drosophila miranda neo-
881 Y chromosome. Genome Biol Evol. 3:1329–1337.
882 Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. 12(4):656–664.
883 Koerich LB, Wang X, Clark AG, Carvalho AB. 2008. Low conservation of gene content in the
884 Drosophila Y chromosome. Nature 456(7224):949–951.
885 Krasovec M, Kazama Y, Ishii K, Abe T, Filatov DA. 2019. Immediate Dosage Compensation Is
886 Triggered by the Deletion of Y-Linked Genes in Silene latifolia. Curr Biol. 29(13):2214–2221.
887 Kurek R, Reugels AM, Glaätzer KH, Bünemann H. 1998. The Y chromosomal fertility factor
888 Threads in Drosophila hydei harbors a functional gene encoding an axonemal dynein $β$ heavy
889 chain protein. Genetics 149(3):1363–1376.
890 Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics
891 34(18):3094–3100.
892 Mahajan S, Bachtrog D. 2017. Convergent evolution of y chromosome gene content in flies. Nat
893 Commun. 8(1):1–13.
894 Mahajan S, Wei KHC, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young
895 Drosophila Y chromosome using single-molecule sequencing and chromatin conformation
896 capture. PLoS Biol. 16(7):e2006348.
897 Mank JE. 2013. Sex chromosome dosage compensation: Definitely not for everyone. Trends
898 Genet. 29(12):677–683.
899 Mcdonald I c, Evenson P, Nickel CA, Johnson OA. 1978. House fly genetics: isolation of a
900 female determining factor on chromosome 4. Ann Entomol Soc Am. 71(5):692–694.
901 McKay DJ, Klusza S, Penke TJR, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW,
902 Tatomer DC, Lieb JD, et al. 2015. Interrogating the function of metazoan histones using
903 engineered gene clusters. Dev Cell. 32(3):373–386.
904 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
905 Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce
906 framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
907 Meisel RP, Gonzales CA, Luu H. 2017. The house fly Y Chromosome is young and minimally
908 differentiated from its ancient X Chromosome partner. Genome Res. 27(8):1417–1426.
909 Meisel RP, Olafson PU, Adhikari K, Guerrero FD, Konganti K, Benoit JB. 2020. Sex
910 Chromosome Evolution in Muscid Flies. G3 (Bethesda) 10(4):1341–1352.
911 Meisel RP, Scott JG. 2018. Using genomic data to study insecticide resistance in the house fly,
912 Musca domestica. Pestic Biochem Physiol. 151:76–81.
913 Montgomery SH, Mank JE. 2016. Inferring regulatory change from gene expression: the
914 confounding effects of tissue scaling. Mol Ecol. 25(20):5114–5128.
915 Neal RM. 2003. Slice sampling. Ann Stat. 31(3):705–767.
916 Orzack SH, Sohn JJ, Kallman KD, Levin SA, Johnston R. 1980. Maintenance of the three sex
917 chromosome polymorphism in the platyfish, Xiphophorus maculatus. Evolution (NY)
918 34(4):663–672.
919 Picard CJ, Johnston JS, Tarone AM. 2012. Genome Sizes of Forensically Relevant Diptera. J
920 Med Entomol. 49(1):192–197.

921 Ponnikas S, Sigeman H, Abbott JK, Hansson B. 2018. Why Do Sex Chromosomes Stop
922 Recombining? Trends Genet. 34(7):492–503.
923 Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R,
924 Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance.
925 Nature 530(7589):228–232.
926 Rand DM, Clark AG, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in
927 Drosophila melanogaster. Genetics 159(1):173–187.
928 Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution (NY)
929 38(4):735–742.
930 Rice WR. 1987. The accumulation of sexually antagonistic genes as a selective agent promoting
931 the evolution of reduced recombination between primitive sex chromosomes. Evolution (NY)
932 41(4):911–914.
933 Rice WR. 1996a. Evolution of the Y sex chromosome in animals. BioScience 46(5):331–343.
934 Rice WR. 1996b. Sexually antagonistic male adaptation triggered by experimental arrest of
935 female evolution. Nature 381(6579):232–234.
936 Roberts RB, Ser JR, Kocher TD. 2009. Sexual conflict resolved by invasion of a novel sex
937 determiner in Lake Malawi cichlid fishes. Science 326(5955):998–1001.
938 Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 17(2):155-
939 158
940 Sardell JM, Cheng C, Dagilis AJ, Ishikawa A, Kitano J, Peichel CL, Kirkpatrick M. 2018. Sex
941 differences in recombination in sticklebacks. G3 (Bethesda) 8(6):1971–1983.
942 Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK,
943 Kasai S, Leichter CA, et al. 2014. Genome of the house fly, Musca domestica L., a global vector
944 of diseases with adaptations to a septic environment. Genome Biol. 15(10):466.
945 Sharma A, Heinze SD, Wu Y, Kohlbrenner T, Morilla I, Brunner C, Wimmer EA, van de Zande
946 L, Robinson MD, Beukeboom LW, et al. 2017. Male sex in houseflies is determined by Mdmd, a
947 paralog of the generic splice factor gene CWC22. Science 356(6338):642–645.
948 Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S,
949 Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is
950 a mosaic of discrete sequence classes. Nature 423(6942):825–837.
951 Soh YQS, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C,
952 Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse y chromosome reveals convergent
953 gene acquisition and amplification on both sex chromosomes. Cell 159(4):800–813.
954 Son JH, Kohlbrenner T, Heinze S, Beukeboom LW, Bopp D, Meisel RP. 2019. Minimal effects
955 of proto-Y chromosomes on house fly gene expression in spite of evidence that selection
956 maintains stable polygenic sex determination. Genetics 213(1):313–327.
957 van Doorn GS, Kirkpatrick M. 2007. Turnover of sex chromosomes induced by sexual conflict.
958 Nature 449(7164):909.
959 van Doorn GS, Kirkpatrick M. 2010. Transitions between male and female heterogamety caused
960 by sex-antagonistic selection. Genetics 186(2):629–645.
961 Vicoso B. 2019. Molecular and evolutionary dynamics of animal sex-chromosome turnover. Nat
962 Ecol Evol. 3:1632–1641.
963 Volff JN. 2006. Turning junk into gold: Domestication of transposable elements and the creation
964 of new genes in eukaryotes. BioEssays 28(9):913–922.
965 Wei KHC, Bachtrog D. 2019. Ancestral male recombination in Drosophila albomicans produced
966 geographically restricted neo-Y chromosome haplotypes varying in age and onset of decay.

967 PLoS Genet. 15(11):e1008502.
968 Zhou Q, Bachtrog D. 2012a. Chromosome-wide gene silencing initiates y degeneration in
969 drosophila. Curr Biol. 22(6):522–525.
970 Zhou Q, Bachtrog D. 2012b. Sex-specific adaptation drives early sex chromosome evolution in
971 Drosophila. Science 337(6092):341–345.
972 Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MTP, Zhang G. 2014.
973 Complex evolutionary trajectories of sex chromosomes across bird taxa. Science
974 346(6215):1246338.

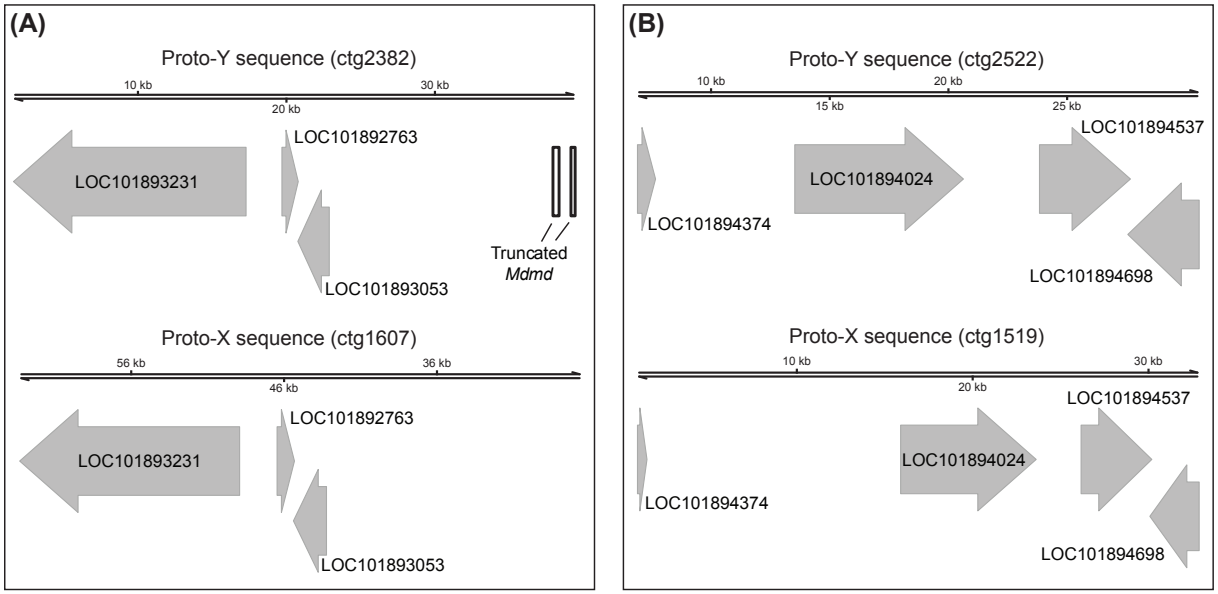
975 **Figure legends**
976
977 **Figure 1**. Elevated heterozygosity on the third and X chromosomes in genotypic (III^M/III) males
978 relative to sex-reversed (III/III) males. The boxplots show the distributions of the percent of
979 heterozygous variants per gene in the genotypic males relative to the sex-reversed males (% III^M
980 heterozygous variants) on each chromosome. Muller element nomenclature for each
981 chromosome is shown in parentheses (Meisel and Scott 2018). See Materials and Methods for
982 the calculation of % III^M heterozygous variants. Values more than 50% indicate more
983 heterozygous variants in genotypic (III^M/III) males, and less than 50% indicates more
984 heterozygous variants in sex-reversed males. The median across autosomes I, II, IV, and V is
985 represented by a dashed line.
986
987 **Figure 2.** Two proto-X and proto-Y loci identified in the III^M genome assembly. (A) One locus
988 was identified with a truncated *Mdmd* on the proto-Y (ctg2382) and the same three genes on both
989 the proto-X and proto-Y. (B) One locus was identified with four genes on both the proto-X and
990 proto-Y. The two contigs (ctg2522 and ctg1519) were assigned to the proto-Y and proto-X based
991 on sequences that are enriched in the male relative to female (see Supplementary Figure 1).
992
993 **Figure 3**. Evidence for moderately elevated ASE on the third (proto-Y) chromosome in III^M
994 males. (A) Proportions of genes with ASE in genotypic (III^M) or sex-reversed (III) males on each
995 chromosome. There is not a significant difference on any chromosome between the two
996 genotypes. (B) Proportions of genes with ASE in genotypic males and non-ASE in sex-reversed
997 males on the third chromosome and all other chromosomes (left two bars). Proportions of genes
998 with non-ASE in genotypic males and ASE in sex-reversed males on the third chromosome and
999 all other chromosomes (right two bars). The asterisk indicates a significant difference ($p<0.05$) in
1000 the number of genes in these categories as determined by Fisher's exact test.
1001
1002 **Figure 4**. Allele-specific expression (ASE) of *Md-HEATR2*. (A) Diagnostic variable sites for
1003 ASE in the *Md-HEATR2* gene are based on haplotypes estimated in IDP-ASE. Read depth is
1004 measured as fragments per million mapped reads (FPM) in III^M males, sex-reversed (SR) males,
1005 and the III^M or III allele in III^M males. (B) Variable sites that differ between genotypic (III^M)
1006 males and sex-reversed males (triangles) across 1,273 base pairs upstream of *Md-HEATR2* were
1007 identified using Oxford Nanopore long reads only. (C) Transcription factor (TF) binding motifs
1008 predicted within 1,273 base pairs upstream of *Md-HEATR2*. The starting position of each motif
1009 is given, along with its length (in parentheses). None of the variable sites overlap with predicted
1010 TF motifs. All positions are coordinates in scaffold NW_004764965.1 of the reference genome
1011 assembly (Scott et al. 2014).

1012 **Figure 1**



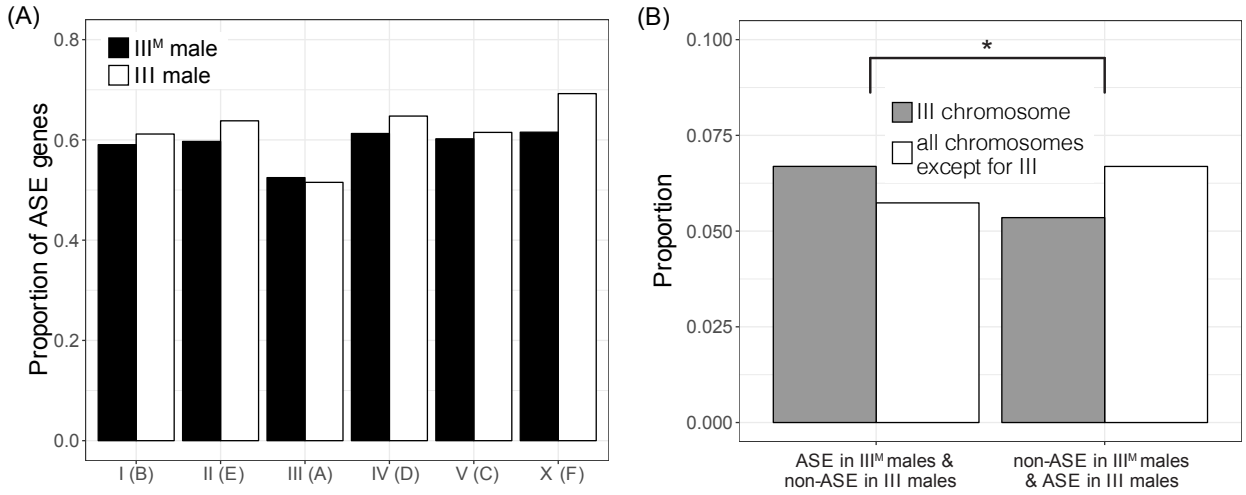
1013

1014 **Figure 2**



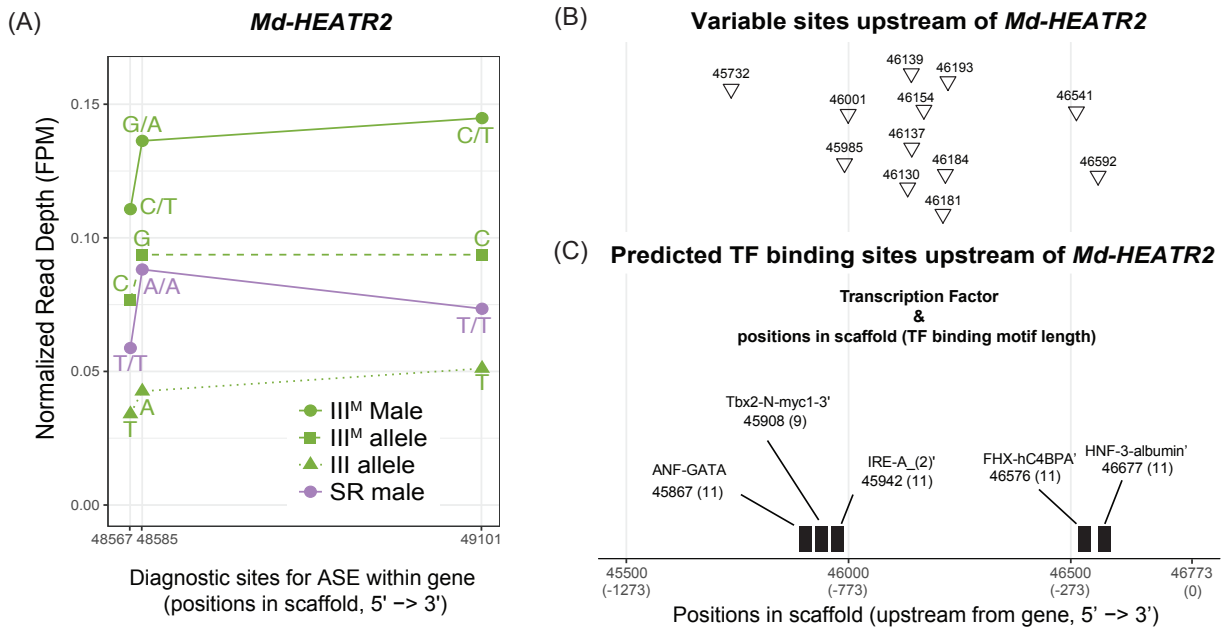
1015

1016 **Figure 3**



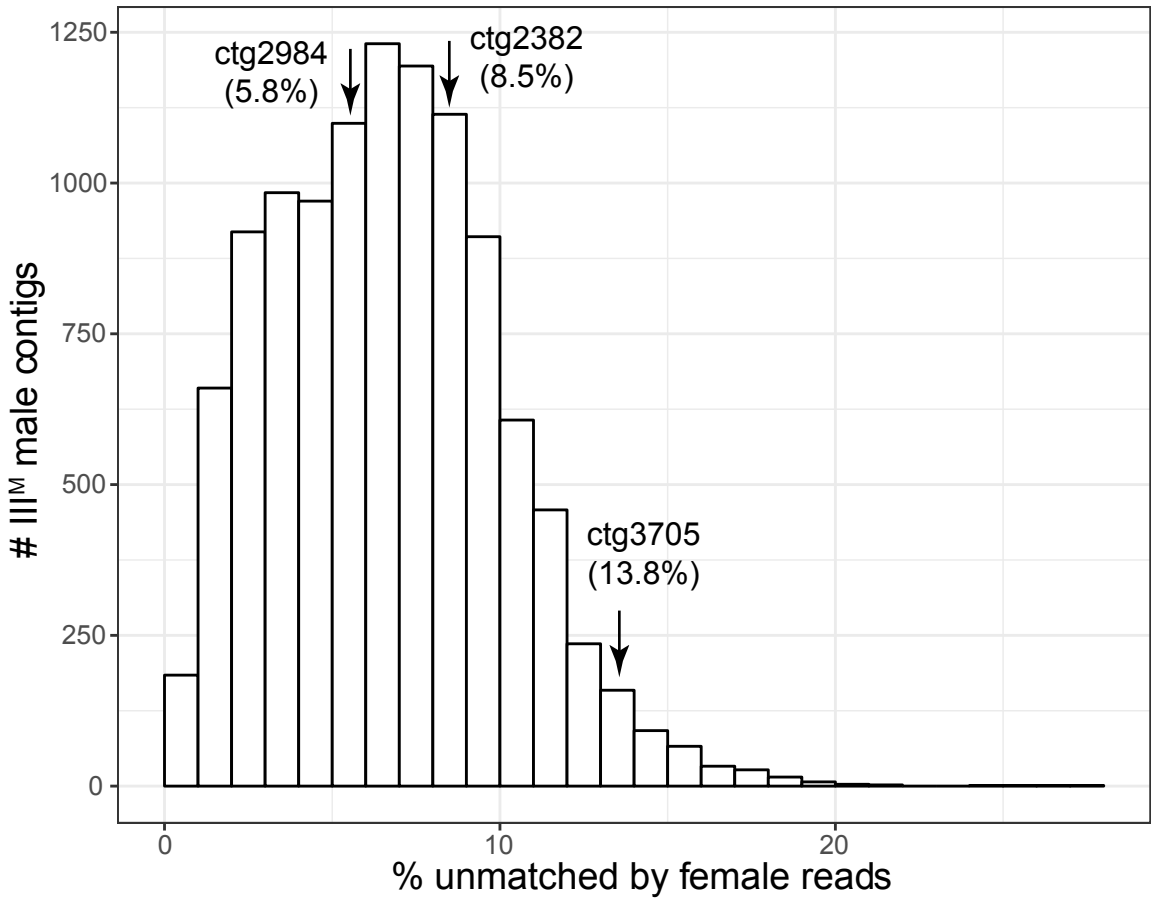
1017

1018 **Figure 4**

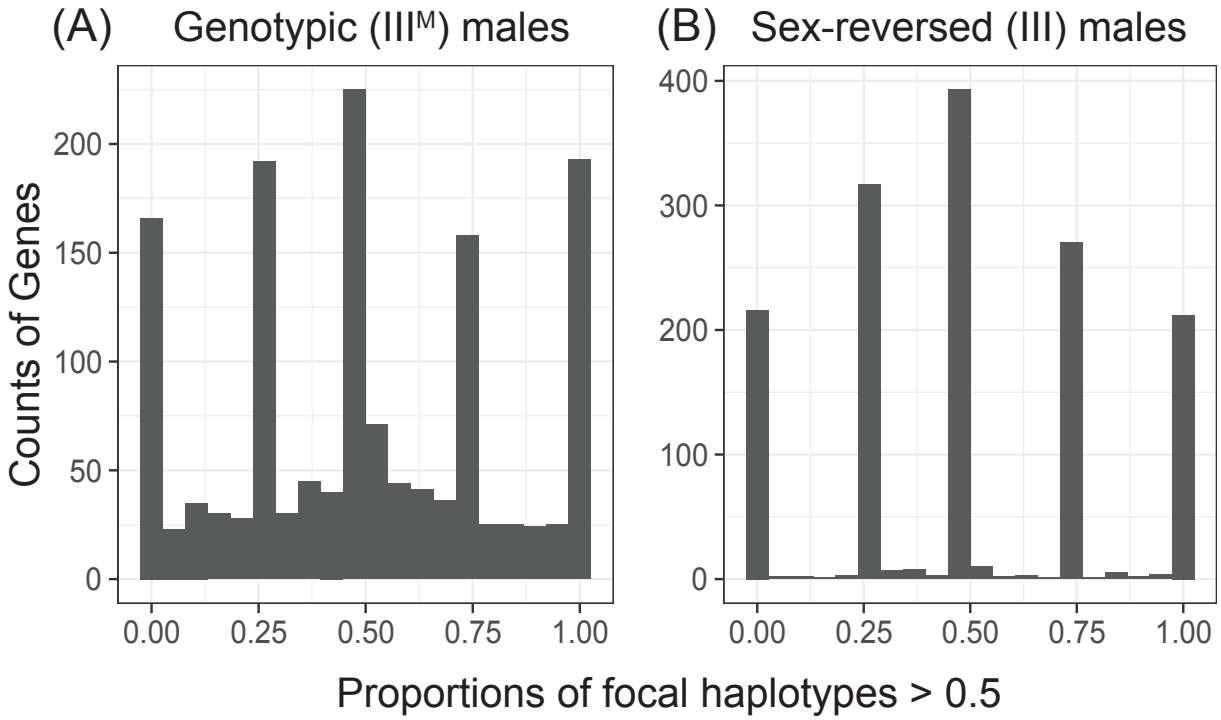


1019

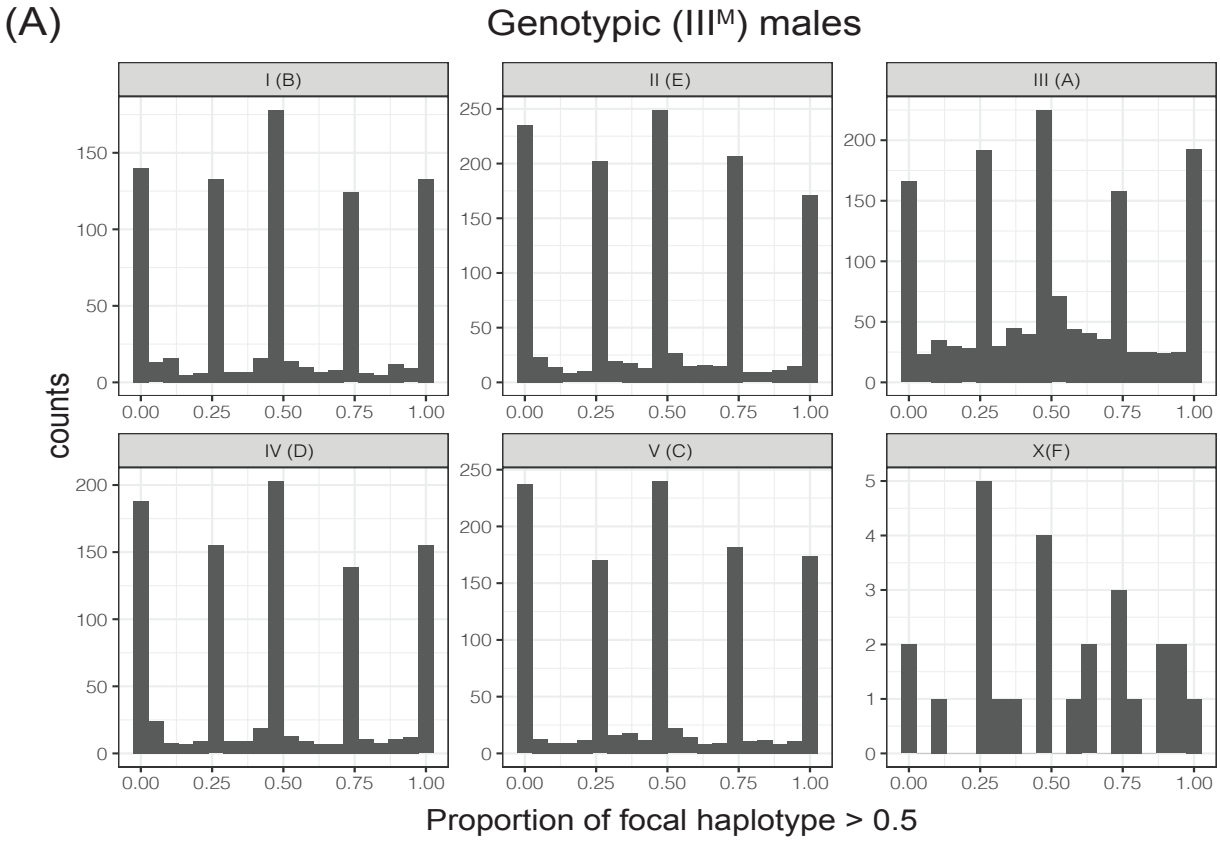
Supplementary Materials

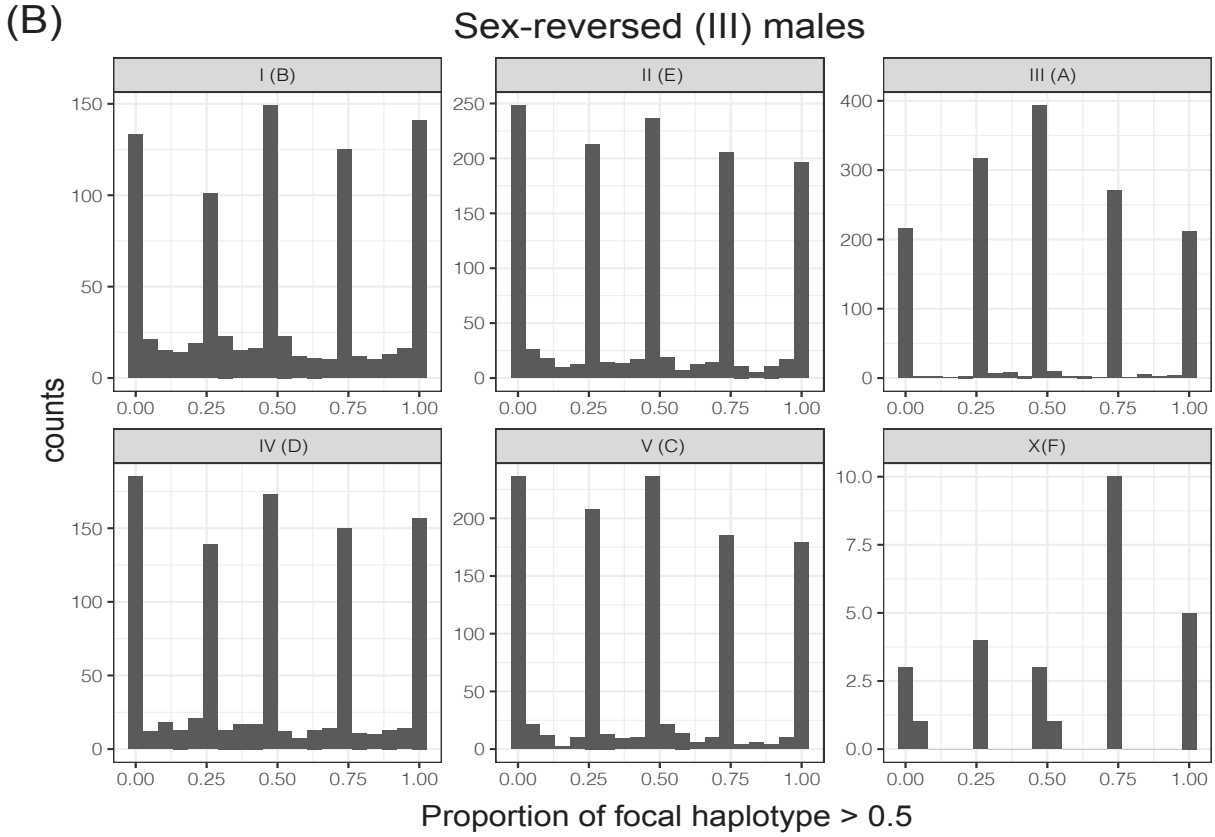


Supplementary Figure 1. Histograms of the percent unmatched by female reads (%UFR) for all contigs in the III^M male genome assembly. Three contigs with *Mdmd* are shown, with their %UFR in parentheses.



Supplementary Figure 2. Histograms of third chromosome genes in different ASE categories. ASE was measured in (A) genotypic (III^M/III) and (B) sex-reversed (III/III) males. ASE is measured as the proportion of iterations in which a focal haplotype is >0.5 of the alleles expressed across iterations in an MCMC simulation. We consider a gene to have ASE if the proportion is <0.125 or >0.875. Genes with a proportion of focal haplotypes between 0.375 and 0.625 are classified as not having ASE.


(A)
Genotypic (III^{M}) males
I (B)
II (E)
III (A)
IV (D)
V (C)
X(F)
counts
Proportion of focal haplotype > 0.5


(B)
Sex-reversed (III) males
I (B)
II (E)
III (A)
IV (D)
V (C)
X(F)
counts
Proportion of focal haplotype > 0.5

Supplementary Figure 3. Histograms of allele-specific expression (ASE) in genes for all chromosome are shown in the genotypic (III^M/III) males and the sex-reversed (III/III) males. Muller element nomenclature (from *Drosophila*) for each chromosome is given in parentheses (Meisel and Scott 2018). If a gene is expressed equally between two alleles, the proportion of the focal haplotype is 0.5; otherwise, the proportion is greater or less than 0.5.

Supplementary Table 1. Nine contigs in the III^M male assembly contain genes that are assigned to the third chromosome and have a %UFR>13.8. The genes on each contig are listed.

Contig ID	Gene	# copies of gene	%UFR of contig
ctg2522	LOC101894374	1	16.9
	LOC101894024	1	
	LOC101894537	1	
	LOC101894698	1	
ctg3539	LOC101888429	1	15.4
	LOC109612838	1	
ctg8407	LOC109613819	1	16.6
ctg1423	LOC101898200	6	14.5
ctg4658	LOC101898200	2	14.0
ctg5526	LOC101898200	1	17.3
ctg8144	LOC101898200	1	14.8
ctg9654	LOC101898200	1	20.1
ctg10432	LOC101898200	1	15.9

Supplementary Table 2. Genes on contigs with differentiated proto-Y (III^M) and proto-X (III) copies. There are two pairs of contigs (each pair contains a proto-Y contig and a proto-X contig). Results of a test for ASE in genotypic (III^M/III) males are shown in the second to last column and ASE of the sex-reversed (III^M/III) males in the last column. ASE was not estimated for some genes due to low sequence coverage.

			ASE category	
Contigs	Gene ID	Protein product	Genotypic (III^M/III) male	Sex-reversed (III/III) male
Proto-Y (III^M): ctg2382 (41,324 bps) Proto-X (III): ctg1607 (66,106 bps)	LOC101893231	probable Xaa-Pro aminopeptidase 3	Moderate ASE	Extreme ASE
	LOC101892763	transcriptional adapter 1	Non-ASE	Extreme ASE
	LOC101893053	delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochodiral	Not estimated	Not estimated
Proto-Y (III^M): ctg2522 (39,000 bps) Proto-X (III): ctg1519 (56,029 bps)	LOC101894374	mitochondrial ornithine transporter 1-like	Not estimated	Not estimated
	LOC101894024	mitochondrial ornithine transporter 1	Moderate ASE	Extreme ASE
	LOC101894537	mitochondrial ornithine transporter 1-like	Extreme ASE	Moderate ASE
	LOC101894698	fast kinase domain-containing protein 5, mitochondrial	Extreme ASE	Moderate ASE

Supplementary Table 3. Missense alleles in coding sequences of two genes on the proto-Y and proto-X contigs. LOC101894698 has also five variable sites that are all synonymous, which are not shown in the table. Variants in coding sequences of the other three genes on the proto-Y and proto-X (LOC101892763, LOC101894024, and LOC101894537) are synonymous and not shown in the table.

Gene	Position at scaffold	III^M male genotype	III male genotype	III^M allele	Reference genome allele
LOC101894698	117354	T/A	A/A	T	T
	118649	C/T	T/T	C	C
	118909	T/C	C/C	T	T
LOC101893231	17382	C/T	C/C	T	C
	17429	A/G	A/A	G	A
	29021	G/A	A/A	G	G

Supplementary Table 4. Counts of ASE genes and non-ASE genes on each chromosome in genotypic (III^M/III) males and sex-reversed (III/III) males.

Chromosome (Muller element)	# genes with ASE in III^M males	# genes with non-ASE in III^M males	# genes with ASE in III males	# genes with non-ASE in III males	Odds ratio	95% CI of odds ratio
III(A)	456	413	438	412	1.038574	0.8555646 - 1.2606663
genome except III(A)	**1635**	**1089**	**1711**	**1010**	**0.886281**	**0.7933539 - 0.9900421**
I(B)	320	222	334	212	0.915002	0.7123784 - 1.1750187
II(E)	465	314	513	291	0.840134	0.6821523 - 1.0344483
IV(D)	394	249	395	215	0.861373	0.6798614 - 1.0908786
V(C)	448	296	460	288	0.947652	0.7654588 - 1.1730543
X(F)	8	8	9	4	0.720539	0.102255 - 4.754933

Supplementary Table 5. Counts of genes with ASE on each chromosome in genotypic (III^M/III) males and sex-reversed (III/III) males. The total number of genes (# genes) in each chromosome group with ASE measurements in both genotypic and sex-reversed males (second column), # genes with ASE in genotypic males and non-ASE in sex-reversed males (third column), and # genes with non-ASE in genotypic males and ASE in sex-reversed males (fourth column) are shown. Bold indicates statistical significance ($P < 0.05$)

		ASE in III^M males and non-ASE in III males	non-ASE in III^M males and ASE in III males	Fisher's Exact Test	
Chromosome (Muller element)	# genes	# genes	# genes	Odds ratio compared with III(A)	95% CI
III(A)	1420	95	76		
genome except III(A)	4201	**241**	**281**	**1.45665**	**1.014895 - 2.095777**
I(B)	824	51	59	1.444139	0.8688985 - 2.4078676
II(E)	1236	**68**	**86**	**1.578592**	**0.9961489 - 2.5100267**
IV(D)	966	**55**	**76**	**1.724145**	**1.063163 - 2.809137**
V(C)	1149	67	77	1.434875	0.8982881 - 2.2981210
X(F)	26	0	0	0	0 - Infinity

Supplementary Table 6. Counts of ASE genes based on the division of ASE measurements into five bins, following the rules described in the Methods. ASE proportions are sorted in the order of extreme (1st and 5th), moderate (2nd and 4th), and no (3rd) ASE. Only extreme ASE was used in the comparisons with non-ASE presented in the main text.

Chr (ME)	Sections for ASE proportions	# of genes in genotypic (III^M/III) Males			# of genes in sex-reversed (III/III) Males		
I(B)	1st (extreme ASE)	167	320	627	167	334	667
	5th (extreme ASE)	153			167		
	2nd (moderate ASE)	158	307		167	333	
	4th (moderate ASE)	149			166		
	3rd (non-ASE)	222	222	222	212	212	212
II(E)	1st (extreme ASE)	270	465	972	291	513	1013
	5th (extreme ASE)	195			222		
	2nd (moderate ASE)	251	507		258	500	
	4th (moderate ASE)	256			242		
	3rd (non-ASE)	314	314	314	291	291	291
III(A)	1st (extreme ASE)	215	456	1043	220	438	1050
	5th (extreme ASE)	241			218		
	2nd (moderate ASE)	312	587		333	612	
	4th (moderate ASE)	275			279		
	3rd (non-ASE)	413	413	413	412	412	412
IV(D)	1st (extreme ASE)	217	394	754	214	395	794
	5th (extreme ASE)	177			181		
	2nd (moderate ASE)	189	360		201	399	
	4th (moderate ASE)	171			198		
	3rd (non-ASE)	249	249	249	215	215	215
V(C)	1st (extreme ASE)	258	448	891	268	460	908
	5th (extreme ASE)	190			192		
	2nd (moderate ASE)	220	443		239	448	
	4th (moderate ASE)	223			209		
	3rd (non-ASE)	296	296	296	288	288	288
X(F)	1st (extreme ASE)	3	8	21	4	9	23
	5th (extreme ASE)	5			5		
	2nd (moderate ASE)	7	13		4	14	
	4th (moderate ASE)	6			10		
	3rd (non-ASE)	5	5	5	4	4	4

Supplementary Table 7. The number of annotated genes assigned to each chromosome (Muller elements in parentheses) and total length of the corresponding chromosomes in base pairs (bp).

Chromosomes (Muller elements)	# genes on chromosomes	Size in length (bp)
I(B)	2128	73015049
II(E)	3175	95424831
III(A)	2271	76855929
IV(D)	2400	76690558
V(C)	2660	81074832
X(F)	47	1486543
Unassigned	3859	345856202
Total	16540	750403944

Supplementary Data 1. A VCF file called with the RNA-seq reads
Supplementary Data 2. A VCF file called with the Oxford Nanopore reads
Supplementary Data 3. IDP-ASE (allele-specific expression) output for the genotypic males
Supplementary Data 4. IDP-ASE (allele-specific expression) output for the sex-reversed males
Supplementary Data 5. k-mer comparison in the III^M male genome assembly

References

Meisel RP, Scott JG. 2018. Using genomic data to study insecticide resistance in the house fly, Musca domestica. Pestic Biochem Physiol. 151:76–81.