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] and Richard P. Meisel[*,1]

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

*Corresponding authors: Jae Hak Son (jaehak.son.225@gmail.com) and 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

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

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

126

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

139

140

141 **Results and Discussion**

142

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

144

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

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

188 contigs containing sequences that are enriched in males relative to females (Carvalho and Clark
189 2013). Our approaches will identify regions on the proto-sex chromosome that contain genes that
190 can be assigned to the third chromosome, are present on both the proto-X and proto-Y with
191 sufficient divergence to assemble separately, and with both the proto-X and proto-Y gametologs
192 assembled in our III^M male genome. We identified two loci, one from each of our two
193 approaches (see Supplementary Materials for details), with sufficient X-Y divergence to
194 assemble into separate proto-X and proto-Y contigs (Figure 2). At each locus, we have one
195 proto-Y contig and one proto-X contig.
196
197 One of the three genes on one proto-Y contig (ctg2382) and all four genes on the other proto-Y
198 contig (ctg2522) are nuclear-encoded mitochondrial genes (Figure 2 and Supplementary Table
199 2). Nuclear-encoded mitochondrial genes can evolve under sexually antagonistic selection (Rand
200 et al. 2001), possibly because of conflicts over mitochondrial functions in sperm and in other
201 tissues (Gemmell et al. 2004; Gallach et al. 2010). These inter-sexual conflicts could favor or
202 select against X-linked mitochondrial genes depending on the extent of co-transmission of the X
203 chromosome and mitochondria in females (Drown et al. 2012; Dean et al. 2014). Our results
204 suggest that mito-nuclear inter-sexual conflicts might also be important for X-Y divergence in
205 young sex chromosomes, where sexually antagonistic variants could be partitioned between
206 female-beneficial X-linked alleles and male-beneficial Y-linked alleles. Additional work is
207 required to test this hypothesis.
208
209 Two of the genes on the proto-Y contigs have differences in their protein-coding sequence from
210 their proto-X chromosome gametologs. One of the two (*LOC101894698*, encoding fast kinase
211 domain-containing protein 5, mitochondrial) contains three missense variable sites at which
212 genotypic (III^M/III) males are heterozygous and sex-reversed (III/III) male are homozygous
213 (Supplementary Table 3). The protein encoded by this gene contains an RNA-binding (RAP)
214 domain, but all three missense variants are not found in the domain. The other gene
215 (*LOC101893231*, which does not have a predicted mitochondrial function) has three missense
216 alleles at which genotypic males are heterozygous and sex-reversed male are homozygous
217 (Supplementary Table 3). One of three missense sites (at position 29,021 in the scaffold of the
218 reference genome) in this gene is found in a proline aminopeptidase P II domain. For all

missense alleles in both genes, we inferred the III allele as the one in common between genotypic and sex-reversed males and the III^M allele as the one unique to genotypic males. However, the inferred III^M alleles are not specific to the III^M chromosome because all III^M alleles in *LOC101894698* and one of the III^M alleles in *LOC101893231* (at position 29,021) were found in the reference genome (which comes from a genotypic female without a III^M chromosome). Therefore, some of the III^M alleles in these genes are segregating as polymorphic variants on the proto-X chromosome.

An alternative explanation of our candidate proto-Y contigs (ctg2382 and ctg2522) is that they are paralogous sequences that have been duplicated on the proto-Y chromosome, creating a second Y-linked copy of the genes within the duplicated region. Intra-chromosomal duplications are a common feature of Y chromosomes (Skaletsky et al. 2003; Hughes et al. 2010; Soh et al. 2014; Bachtrog et al. 2019; Ellison and Bachtrog 2019). It is therefore possible that the sequences we classify as on the proto-X are actually present on both the proto-X and proto-Y, while the contigs we classify as proto-Y sequences are intra-chromosomal duplications on the proto-Y. However, even in this scenario, the proto-Y sequences are still unique to the proto-Y. Therefore, our interpretations are unlikely to be affected by whether the sequences we classify as proto-Y are true gametologs of the proto-X or intra-chromosomal duplications on the proto-Y.

Gene expression divergence between the proto-Y and proto-X chromosomes

We next tested if differences in *cis*-regulatory sequences between the proto-Y and proto-X chromosomes contribute to expression differentiation. We quantified differential expression between the proto-X and proto-Y chromosome copies of the third chromosome genes by measuring allele-specific expression (ASE) in normal (genotypic) males carrying a III^M proto-Y chromosome and sex-reversed (III/III) males with no proto-Y chromosome. Comparing genotypic and sex-reversed males allows us to control for the effect of sexually dimorphic gene expression on the inference of divergence between the proto-Y (III^M) and proto-X (III) chromosomes.

249 To quantify ASE of genes in genotypic (III^M/III) and sex-reversed (III/III) males, we used
250 existing RNA-seq data (Son et al. 2019) along with our new Oxford Nanopore long read
251 sequencing data. We used the IDP-ASE pipeline (Deonovic et al. 2017), which is more accurate
252 when a hybrid of short and long reads is provided as input. This is because the long reads are
253 used to construct haplotypes, which are used to estimate ASE from the RNA-seq data (see
254 Materials and Methods). This approach differs from our comparison of proto-X and proto-Y
255 contigs because the ASE analysis does not require separate assembly of proto-X and proto-Y
256 contigs. Instead, IDP-ASE uses the raw long read sequences from genotypic and sex reversed
257 males to phase haplotypes when inferring ASE. We measured ASE as the proportion of iterations
258 in a Markov chain Monte Carlo (MCMC) simulation in which the expression of a focal
259 haplotype is estimated as >0.5. This proportion gives a measure of ASE ranging from 0 (extreme
260 ASE in favor of one allele) to 1 (extreme ASE in favor of another allele), with 0.5 indicating
261 equal expression of both alleles. We are unable to determine if the focal haplotype refers to the
262 III^M or III allele across the entire third chromosome, but we do differentiate between these alleles
263 for a handful of genes where we are able to perform manual curation (see below).
264
265 We assigned each gene with sufficient expression data into one of five bins of ASE. The
266 proportions of iterations with focal haplotypes >0.5 were overrepresented at five values (0, 0.25,
267 0.5, 0.75, and 1) both in genotypic and sex-reversed males (Supplementary Figure 2 and 3).
268 These proportions may be overrepresented because we only sampled two genotypes for our ASE
269 analysis, which caused us to have a non-continuous distribution of proportions. We divided the
270 proportion of iterations with the focal haplotypes >0.5 into five bins, with each bin capturing one
271 of the five most common proportions (Supplementary Figure 2 and 3): 1) extreme ASE, with a
272 value between 0 and 0.125; 2) moderate ASE, with a value between 0.125 and 0.375; 3) non-
273 ASE, with a value between 0.375 and 0.625; 4) moderate ASE, with a value between 0.625 and
274 0.875; and 5) extreme ASE, with a value between 0.875 and 1. In the analysis below, we
275 considered a gene to have ASE if it falls into one of the two bins of extreme ASE; genes in the
276 non-ASE bin (proportion of focal haplotype >0.5 between 0.375 and 0.625) were classified as
277 having non-ASE. Genes with moderate ASE were excluded from most of our analyses in order to
278 be conservative about ASE assignment.
279

280 We first investigated ASE of genes we identified at the two loci where we have separate proto-X
281 and proto-Y contigs from our genome assembly (Supplementary Table 2). We had enough
282 sequence coverage and heterozygous sites to estimate ASE in 5/7 genes. Two mitochondrial
283 genes (*LOC101894537* and *LOC101894698*) had extreme ASE in genotypic males and moderate
284 ASE in sex-reversed males. The elevated ASE in genotypic males is suggestive of expression
285 divergence between the proto-Y and proto-X copies. Three genes (*LOC101892763*,
286 *LOC101893231*, and *LOC101894024*) exhibited extreme ASE in sex-reversed males and non-
287 ASE or moderate ASE in genotypic males. Elevated ASE in sex-reversed males is not expected
288 because they are genotypic females with two copies of the proto-X chromosome. This result
289 suggests there is not large-scale expression divergence between the proto-Y and proto-X
290 chromosomes. The limited expression divergence (Supplementary Table 2) and paucity of fixed
291 amino acid differences (Supplementary Table 3) between proto-Y and proto-X gametologs are
292 consistent with minimal differentiation between the house fly proto-Y and proto-X
293 chromosomes.
294
295 We next examined ASE across the entire third chromosome. If the III^M proto-Y chromosome is
296 differentiated in gene expression from its homologous III proto-X chromosome because of
297 differences in *cis*-regulatory alleles across the entire third chromosome, then we expect a higher
298 fraction of genes with ASE on the third chromosome in the genotypic (III^M/III) males than in the
299 sex-reversed (III/III) males. In contrast to that expectation, we did not find an excess of genes
300 with ASE in genotypic males compared to ASE genes in sex-reversed males on the third
301 chromosome relative to other chromosomes (Figure 3A and Supplementary Table 4; Fisher's
302 exact test, $P = 0.6996$). This result suggests that the III^M proto-Y chromosome is not broadly
303 differentiated in *cis*-regulatory alleles from the standard third (proto-X) chromosome. This
304 provides evidence that the early stages of Y chromosome evolution do not involve chromosome-
305 wide changes in gene regulation.
306
307 We next identified individual genes with differences in ASE between genotypic (III^M/III) and
308 sex-reversed (III/III) males. There are 95 third chromosome genes with ASE in the genotypic
309 males that are non-ASE in the sex-reversed males (Supplementary Table 5). These genes could
310 have ASE in III^M males because of differences in *cis* regulatory sequences between the III^M and

311 standard third chromosome. To test whether the observed number of third chromosome genes
312 with ASE in genotypic males that are non-ASE in sex-reversed males is in excess of a null
313 expectation, we determined the number of third chromosome genes with ASE in sex-reversed
314 males that are non-ASE in genotypic males (i.e., the opposite of what we did above to find the
315 first set of 95 genes). There are 76 third chromosome genes with ASE in the sex-reversed males
316 that are non-ASE in genotypic males (Supplementary Table 5). We also identified 241 genes on
317 other chromosomes with ASE in genotypic males that are non-ASE in sex-reversed males, as
318 well as 281 genes on other chromosomes with ASE in sex-reversed males that are non-ASE in
319 genotypic males (Figure 3B and Supplementary Table 5). We do not expect any difference in
320 ASE between genotypic and sex reversed males for chromosomes other than the third.
321 Comparing genes with and without ASE on the third chromosome and the rest of the genome,
322 there is indeed an excess of third chromosome genes with ASE in genotypic males that are non-
323 ASE in sex-reversed males (Figure 3B and Supplementary Table 5; Fisher's exact test, P =
324 0.03467). These results suggest that, while the III^M proto-Y chromosome is not broadly
325 differentiated in *cis*-regulatory sequences from the standard third (proto-X) chromosome, there is
326 an excess of individual genes with *cis*-regulatory differences between the III^M proto-Y and its
327 homologous proto-X chromosome.
328
329 Male-specific selection on individual proto-Y genes could be responsible for ASE in genotypic
330 III^M males. In this scenario, male-specific selection would favor *cis*-regulatory variants that drive
331 up- or down-regulation of the III^M copy of a gene (Parsch and Ellegren 2013; Mank 2017). These
332 sex-specific selection pressures are expected to have the greatest effect for genes closest to the
333 male-determining *Mdmd* locus because those genes are most likely to be co-inherited with *Mdmd*
334 (Charlesworth et al. 2014). Unfortunately, we lack a chromosome-scale assembly of the house
335 fly genome (Scott et al. 2014; Meisel and Scott 2018), which prevents us from testing if genes
336 with ASE are clustered on the third chromosome in close proximity to the *Mdmd* locus.
337
338 **Up-regulation of the Y-allele and male-biased expression of a testis-expressed gene**
339
340 We next tested if genes with ASE are differentially expressed between genotypic males, sex-
341 reversed males, and females. A relationship between ASE and sexually dimorphic expression

342 would suggest that up- or down-regulation of Y-linked alleles affects sexually dimorphic
343 phenotypes. We started by selecting genes that we had previously identified as differentially
344 expressed between genotypic (III^M/III) and sex-reversed (III/III) males (Son et al. 2019). These
345 two genotypes have nearly the same expression profiles, with a small number of differentially
346 expressed genes. We were specifically interested in genes on the third (proto-sex) chromosome
347 with "discordant sex-biased expression". Discordant sex-biased genes have male-biased
348 expression in genotypic males (i.e., upregulated relative to phenotypic females) and female-
349 biased expression in the sex-reversed males (downregulated relative to phenotypic females), or
350 vice versa. This pattern of expression is suggestive of *cis*-regulatory divergence between the
351 proto-Y and proto-X chromosomes—a hypothesis that was presented but not tested in our
352 previous study (Son et al. 2019). Here, we test that hypothesis by determining if any third
353 chromosome genes with discordant sex-biased expression have ASE consistent with *cis*-
354 regulatory divergence between the proto-Y (III^M) and proto-X (III) alleles.
355
356 We identified a single gene (*LOC101899975*) with discordant sex-biased gene expression out of
357 the 95 genes on the third chromosome with ASE in the genotypic (III^M/III) males that are non-
358 ASE in the sex-reversed (III/III) males. This gene is homologous to *dynein assembly factor 5,*
359 *axonemal* (human gene *DNAAF5* and *Drosophila melanogaster* gene *HEATR2*). The gene, which
360 we refer to as *M. domestica HEATR2* (*Md-HEATR2*), is expected to encode a protein that
361 functions in flagellated sperm motility (Diggle et al. 2014), and it has strong testis-biased
362 expression in *D. melanogaster* (Chintapalli et al. 2007). *Md-HEATR2* has male-biased expression
363 in the abdomens of genotypic males and female-biased expression in the abdomens of sex-
364 reversed males (Son et al. 2019), suggesting that expression differences between the III^M proto-Y
365 and the standard third (proto-X) chromosome cause the male-biased expression of the gene in the
366 genotypic males.
367
368 We identified three diagnostic variant sites for ASE within *Md-HEATR2* (Figure 4A), which are
369 all synonymous SNPs. The genotypic (III^M/III) males are heterozygous and the sex-reversed
370 (III/III) males are homozygous at all diagnostic sites. We inferred the allele on the standard third
371 chromosome as the one in common between genotypic and sex-reversed males, and the III^M
372 allele as the one unique to genotypic males at each diagnostic variant site. Curiously, all three

373 III^M alleles are found in the reference genome, suggesting that these synonymous variants are not
374 fixed differences between the proto-Y and proto-X. *Md-HEATR2* is expressed higher in III^M
375 genotypic males than in sex-reversed males (Figure 4A). In the III^M genotypic males, the III^M
376 (Y-linked) alleles are expressed higher than the X-linked alleles, indicating that the Y-linked
377 alleles are associated with the up-regulation of the gene in III^M genotypic males relative to sex-
378 reversed males (Figure 4A). The copy of *Md-HEATR2* on the III^M proto-Y chromosome is
379 therefore up-regulated relative to the proto-X copy, consistent with higher expression of *Md-*
380 *HEATR2* in genotypic males.
381
382 Using our Nanopore sequencing reads mapped to the reference genome, we examined 1,273 base
383 pairs upstream of *Md-HEATR2* to identify diagnostic sites that could be responsible for
384 regulating the expression differences between the proto-X and proto-Y alleles. We chose that
385 distance because it includes the first variable site we could identify on the scaffold containing
386 *Md-HEATR2* in our Nanopore data (i.e., including a larger region would not provide any
387 additional information). We found twelve variable sites with different alleles (SNPs and small
388 indels) between genotypic (III^M/III) and sex-reversed (III/III) males (Figure 4B). We next
389 examined whether these sites are located within a potential transcription factor (TF) binding
390 region. We found five TF binding regions predicted upstream of *Md-HEATR2* using the
391 'Tfsitescan' tool in the 'object-oriented Transcription Factors Database' (Ghosh 2000). However,
392 none of the twelve variable sites are found within any predicted TF binding regions (Figure 4C).
393 It is possible that the *cis*-regulatory sequences responsible for differential expression are located
394 outside of the region we were able to investigate, which would be consistent with our failure to
395 identify fixed differences in the exons between the proto-Y and proto-X copies of *Md-HEATR2*.
396 A long distance would reduce the genetic linkage between the *cis*-regulatory region and
397 transcribed sequence, allowing for the X-Y differentiation of the regulatory region without
398 differentiation of the transcribed gene. Further work is needed to determine how the differential
399 expression of the proto-X and proto-Y copies of *Md-HEATR2* is regulated.
400
401 We hypothesize that the upregulation of the proto-Y copy of *Md-HEATR2* is the result of
402 selection for higher expression in males. We think that the *cis*-regulatory region is under
403 selection, rather than the protein-coding sequence, because we fail to find any protein coding

404 differences between the proto-X and proto-Y copies, and the synonymous differences we
405 observe are not fixed differences. HEATR2 is involved in dynein arm assembly (Diggle et al.
406 2014), and axonemal dynein is essential for flagellated sperm motility (Kurek et al. 1998;
407 Carvalho et al. 2000). Therefore, it may be beneficial to male fitness to have higher expression of
408 *Md-HEATR2* in testis. However, HEATR2 also functions in mechanosensory neurons in
409 *Drosophila* (Diggle et al. 2014). There may be conflict over the *cis*-regulatory sequences that
410 promote expression in testis and neurons, which may prevent up-regulation of the proto-X copy
411 of *Md-HEATR2* in testis. These opposing (i.e., sexually antagonistic) selection pressures on *Md-*
412 *HEATR2* expression would be resolved in a Y-linked copy that is only under selection in males
413 (Rice 1996b). Alternatively, the expression differences between sex-reversed and genotypic
414 males could be caused by differences in tissue scaling between genotypes (Montgomery and
415 Mank 2016). For example, if genotypic males have larger testes and if the proto-Y allele is
416 preferentially expressed in testis, then it will appear as if genotypic males have higher expression
417 of *Md-HEATR2*. Even in this model, there is ASE associated with testis-specific expression,
418 which is consistent with male-specific selection favoring the proto-Y allele.
419
420 The up-regulation or testis-biased expression of the proto-Y copy of *Md-HEATR2* may not
421 completely resolve sexually antagonistic selection pressures on *Md-HEATR2* expression because
422 of three notable features of house fly genetics. First, Y-linked alleles are only under male-
423 specific selection when they are in complete genetic linkage with a male-determining gene
424 (Charlesworth et al. 2014). Complete linkage between loci can be caused by recombination
425 suppressors (e.g., chromosomal inversions) that prevent genetic exchange between the X and Y
426 chromosomes (Rice 1987; Charlesworth 2017). While there is no direct evidence for inversions
427 or other suppressors of recombination on the house fly third chromosome, crossing over in male
428 meiosis is rare in most flies (Gethmann 1988). A general lack of recombination in males would
429 prevent X-Y exchange. However, male recombination has been documented in house fly
430 (Feldmeyer et al. 2010), suggesting that X-Y exchange is possible. Second, female house flies
431 can carry proto-Y chromosomes if they also carry the female-determining *Md-tra*^{D} allele (Hamm
432 et al. 2015). Female-transmission of the proto-Y provides another avenue for X-Y exchange in
433 the absence of an inversion or other suppressor of recombination on the proto-Y or proto-X. We
434 observe alleles in *Md-HEATR2* and other third chromosome genes that are found on both the

435 proto-Y and proto-X (Supplementary Table 3), which is consistent with either X-Y
436 recombination or ancestral polymorphisms that predate the formation of the proto-sex
437 chromosome. Third, when females carry a proto-Y chromosome, there will be selection against
438 male-beneficial, female-detrimental sexually antagonistic alleles, which will further prevent the
439 resolution of sexual conflict.
440
441 Selection on proto-Y alleles in females, along with X-Y recombination that moves proto-X
442 alleles onto the proto-Y, will reduce the efficacy of male-specific selection pressures on sexually
443 antagonistic alleles of *Md-HEATR2*. The extent of these effects will depend on the frequency of
444 proto-Y chromosomes found in females. First, the extent to which selection in females acts
445 against male-beneficial sexually antagonistic alleles depends on the frequency with which the
446 proto-Y chromosomes are found in females (Rice 1984). Second, if recombination rate is
447 sexually dimorphic, the rate of population level recombination between the male-determiner and
448 *Md-HEATR2* will depend on how often the proto-Y chromosomes are found in females (Sardell
449 et al. 2018). Notably, the frequencies of proto-Y chromosomes and the female-determining *Md-*
450 tra^D allele vary across populations (Hamm et al. 2015; Meisel et al. 2016). Therefore, the
451 frequency with which females carry proto-Y chromosomes will vary across populations,
452 suggesting that selection against male-beneficial *cis*-regulatory alleles of *Md-HEATR2* on the
453 proto-Y may be population-specific.
454
455 The expression divergence of the proto-X and proto-Y copies of *Md-HEATR2* could constitute
456 an early stage of X-Y differentiation before chromosome-wide X-Y differentiation occurs
457 (Bachtrog 2013). Young Y chromosomes have very similar gene content as their ancestral
458 autosomes. In contrast, old Y chromosomes are more likely to have only retained genes with
459 male-specific functions or recruited genes associated with testis expression from other autosomes
460 (Koerich et al. 2008; Kaiser et al. 2011; Mahajan and Bachtrog 2017). Our results suggest that
461 changes in the expression of individual Y-linked genes that were retained from the ancestral
462 autosome could have important phenotypic effects during early Y chromosome evolution,
463 consistent with gene-by-gene divergence on the young Y chromosomes (Wei and Bachtrog
464 2019).
465

466 **Conclusions**

467

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

480

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

497 third chromosome only in genotypic males, and not in sex-reversed males, is small; we identify
498 fewer than 100 genes that meet these criteria, which is less than 5% of the 2,524 genes assigned
499 to the third chromosome (Meisel and Scott 2018).
500
501 We identified one gene on the third chromosome (*Md-HEATR2*) with ASE in genotypic males
502 that is non-ASE in sex-reversed males and has discordant sex-biased expression between
503 genotypic and sex-reversed males (Figure 4). We hypothesize that expression divergence of *Md-*
504 *HEATR2* could be an example of very early X-Y differentiation of individual genes that results
505 from sex-specific selection. Notably, *Md-HEATR2* is expected to have important functions in
506 spermatogenesis or sperm motility (Diggle et al. 2014; Fuller 1993), similar to the five
507 mitochondrial genes contained within genomic regions that are divergent between the proto-Y
508 and proto-X. Male fertility is an important target of selection during the evolution of old Y
509 chromosomes (Carvalho et al. 2009; Hughes et al. 2010), and our results suggest that selection
510 on male fertility is also an important driver of X-Y divergence during the early evolution of sex
511 chromosomes. Our results also suggest that these selection pressures during the earliest stages of
512 Y chromosome evolution drive gene-by-gene, rather than chromosome-scale, changes in gene
513 expression.
514
515
516 **Materials and Methods**
517
518 **<u>Fly strain</u>**
519
520 We analyzed RNA-seq data and performed Oxford Nanopore sequencing on a house fly strain
521 that allows for identification of genotypic (III^M/III) males and sex-reversed (III/III) males
522 (Hediger et al. 2010). This is because the standard third chromosome (III) in this strain has the
523 recessive mutations *pointed wing* and *brown body*. Sex-reversed males (and normal females)
524 have both mutant phenotypes, whereas genotypic males are wild-type for both phenotypes
525 because the III^M chromosome has the dominant wild-type alleles. The RNA-seq data that we
526 analyzed (available at NCBI GEO accession GSE126689) comes from a previous study that used
527 double-stranded RNA (dsRNA) targeting *Md-tra* to create sex-reversed phenotypic males that

528 have a female genotype without a proto-Y chromosome (Son et al. 2019). This is because active
529 *Md-tra* drives female development and inactive *Md-tra* triggers male development (Hediger et
530 al. 2010). We compared gene expression in the abdomens of sex-reversed males, genotypic
531 males that received a sham treatment of dsRNA targeting GFP, and genotypic females that are
532 phenotypically female. We analyzed data from three replicates of each genotype and treatment,
533 with each replicate consisting of a single fly abdomen (Son et al. 2019). We used genotypic
534 males and sex-reversed males from the same strain that were subjected to the same treatment for
535 genome sequencing using the Oxford Nanopore long read technology.
536
537 **Oxford Nanopore sequencing**
538
539 We performed Oxford Nanopore sequencing of one genotypic (III^M/III) male and one sex-
540 reversed (III/III) male created from the same strain and using the same *Md-tra* dsRNA treatment
541 as a previous RNA-seq study (Son et al. 2019). DNA was isolated with a phenol/chloroform
542 protocol (see Supplementary Materials). Oxford Nanopore Sequencing libraries were prepared
543 with the 1D genomic DNA Ligation kit (SQK-LSK109, Oxford Nanopore), following the
544 manufacturer's protocol. DNA from the genotypic male and sex-reversed male was used to
545 create a separate sequencing library for each genotype. Following the manufacturer's protocol,
546 15 uL of each library, along with sequencing buffer and loading beads (totaling 75 uL), were
547 separately loaded onto two different R9.4 flow cells (i.e., the libraries from the genotypic and
548 sex-reversed males were run on separate flow cells) until no pores were available on a MinION
549 sequencer (Oxford Nanopore).
550
551 We performed two different base calling pipelines, using the Guppy pipeline software version
552 3.1.5 (Oxford Nanopore). First, for the IDP-ASE analysis, we used parameter options with "--
553 calib_detect --qscore_filtering --min_qscore 10". Second, for the genome assembly we used the
554 default parameters. We used different parameters for IDP-ASE because base quality affects
555 accurate haplotyping used to estimate ASE (see below), and we therefore used a higher threshold
556 for the base quality. For the IDP-ASE analysis, the base called reads were aligned to the house
557 fly genome assembly v2.0.2 (Scott et al. 2014) using Minimap2 version 2.17 with the "-ax map-
558 ont" parameter (Li 2018).

559

560 **Genome assembly, transcript alignment, and sequence divergence**

561

562 We used wtdbg2 to assemble our base called Oxford Nanopore reads using the default
563 parameters (Ruan and Li 2020). To find genic regions in our genome assembly, we aligned
564 house fly transcripts (from Annotation Release 102) as a query against our genome assembly
565 contigs using BLAT with a minimum mapping score of 50 (Kent 2002). We considered BLAT
566 alignments to be gene copies only if the matched sequence length in the BLAT alignments
567 covers at least half of the original (annotated) transcript length. We selected contigs in our
568 assembly that contain genes that are assigned to the third chromosome in the reference genome.

569

570 We used two approaches to differentiate the proto-Y and proto-X contigs. In both approaches we
571 specifically focused on pairs of contigs that contain the same genes because they are indicative
572 of X-Y divergence (i.e., one contig likely contains a proto-Y sequence, and the other contains a
573 proto-X sequence). First, we tested if one contig contains a copy of *Mdmd*, which allows us to
574 designate that contig as the proto-Y copy (and the other is proto-X). *Mdmd* is not present in the
575 annotated house fly genome because the genome was sequenced from female DNA (Scott et al.
576 2014). To find *Mdmd* copies in our genome assembly, we used the sequence of the *Mdmd*
577 transcript (Sharma et al. 2017) as a query to find contigs containing *Mdmd* with BLAST using an
578 e-value cutoff of 1e-3 (Altschul et al. 1990). Second, we used a *k*-mer comparison approach to
579 differentiate the proto-X and proto-Y contigs (Carvalho and Clark 2013). In this approach, we
580 used a *k*-mer size of 15 to measure the percent unmatched by female reads (%UFR) for every
581 contig in our assembly, with higher values indicating an increased likelihood that a contig is Y-
582 linked. The female reads were from the genome project (BioProject accession PRJNA176013;
583 Scott et al. 2014), and we included the III^M male Oxford Nanopore sequencing reads we
584 generated for validation of the bit-array. We followed the options suggested to identify Y
585 sequences in *Drosophila* genomes (Carvalho and Clark 2013), as described previously (Meisel et
586 al. 2017). We used the R package Gviz (Hahne and Ivanek 2016) to visualize genes that are
587 present in III^M (proto-Y) contigs and III (proto-X) contigs.

588

589 **Variant calling**

590

591 We used available RNA-seq data (Son et al. 2019) to identify genetic variants (SNPs and small
592 indels) that differentiate the III^M proto-Y chromosome from the standard third (proto-X)
593 chromosome, and then we tested if III^M males have elevated heterozygosity on the third
594 chromosome as compared to sex-reversed males (Meisel et al. 2017). We used the Genome
595 Analysis Toolkit (GATK) pipeline for calling variants in the RNA-seq data from the *Md-tra*
596 RNAi experiment in (Son et al. 2019), following the best practices for SNP and indel calling on
597 RNA-seq data (McKenna et al. 2010; Meisel et al. 2017). First, we used STAR (Dobin et al.
598 2013) to align reads from three genotypic (III^M/III) male libraries and three sex-reversed (III/III)
599 male libraries to the (female) reference assembly v2.0.2 (Scott et al. 2014). The reference
600 genome was sequenced from the aabys strain, which differs from the one we used in our
601 experiment and has males that carry the Y^M proto-Y chromosome. The aligned reads were used
602 to generate a new reference genome index from the detected splice junctions in the first
603 alignment run, and then a second alignment was performed with the new reference. We next
604 marked duplicate reads from the same RNA molecule and used the GATK tool
605 'SplitNCigarReads' to reassign mapping qualities to 60 with the 'ReassignOneMappingQuality'
606 read filter for alignments with a mapping quality of 255. Indels were detected and realigned with
607 'RealignerTargetCreator' and 'IndelRealigner'. The realigned reads were used for base
608 recalibration with 'BaseRecalibrator' and 'PrintReads'. The base recalibration was performed in
609 three sequential iterations in which recalibrated and filtered reads were used to train the next
610 round of base recalibration, at which point there were no beneficial effects of additional base
611 recalibration as verified by 'AnalyzeCovariates'. We next used the recalibrated reads from all
612 three replicates of genotypic and sex-reversed males to call variants using 'HaplotypeCaller'
613 with emission and calling confidence thresholds of 20. We applied 'genotypeGVCFs' to the
614 variant calls from the two types of males for joint genotyping, and then we filtered the variants
615 using 'VariantFiltration' with a cluster window size of 35 bp, cluster size of 3 SNPs, FS > 20,
616 and QD < 2. The QD filter simultaneously considers read depth and variant quality so that a
617 separate read depth filter is not needed during variant filtration. Because this filtration is applied
618 during the joint genotypic step, all variants are genotyped using the combined information across
619 both types of males, eliminating the need to separately cross-reference read mapping information
620 across genotypes. The final variant calls were used to identify heterozygous variants within

621 genes and to estimate ASE with the IDP-ASE tool (see below) using the coordinates from the
622 genome sequencing project, annotation release 102 (Scott et al. 2014). If a locus has different
623 variants (e.g. a reference allele and an alternative allele), we considered the locus heterozygous.
624 We measured relative heterozygosity within each gene in genotypic (III^M/III) and sex-reversed
625 (III/III) males as the number of heterozygous variants in genotypic males for a given gene (h_G)
626 divided by the total number heterozygous variants in both genotypic and sex-reversed males
627 (h_{SR}), times one hundred: $100h_G/(h_G + h_{SR})$. To annotate each variant within coding sequences,
628 we used SnpEff with filter options "-onlyProtein", "-no-intergenic", "-no-downstream", "-no-
629 upstream", "-no-intron", and "-no-utr" (Cingolani et al. 2012).
630
631 For the variant calling from Nanopore long reads, the base called reads were indexed using fast5
632 files with the 'index' module of Nanopolish version 0.11.1 (Quick et al. 2016), and they were
633 aligned with Minimap2 version 2.17 (Li 2018) to house fly genome assembly v2.0.2 (Scott et al.
634 2014). The aligned and raw reads were used to call variants using the "variants" module of
635 Nanopolish version 0.11.1 with the "--ploidy 2" parameter (Quick et al. 2016). We used a python
636 script 'nanopolish_makerange.py' provided in the package to split the genome into 50 kb
637 segments because it was recommended to use the script for large datasets with genome size more
638 than 50 kb.
639
640 **Allele-specific expression**
641
642 ASE is the unequal expression of the maternal and paternal alleles in a diploid. Estimating ASE
643 with a single reference genome generates bias in the ASE measurement because RNA-seq reads
644 from the allele found in the reference genome could preferentially map to the reference genome
645 relative to reads from alternative alleles (Stevenson et al. 2013). This read-mapping bias can be
646 reduced by filtering out clusters of variants found in close proximity (Stevenson et al. 2013;
647 Zimmer et al. 2016). We accomplished this by excluding all clusters of 3 or more SNPs found
648 within windows of 35 bp (see above), which is similar to the recommended filtering parameters
649 to reduce mapping bias (Stevenson et al. 2013). Our filtering step retained only 21.3% of all
650 SNPs, which we consider to be high confidence SNPs for inferring ASE. In comparison, Zimmer
651 et al. (2016) applied a filter of 6 SNPs per 100 bp, which we find retains 30.4% of all SNPs in

652 our data. Therefore, the filter we have applied to remove SNP clusters is as stringent or more
653 stringent than those previously used to reduce mapping biases that could affect measurements of
654 ASE.
655
656 We investigated if there is elevated ASE on the third chromosome in males carrying one III^M
657 proto-Y and one proto-X chromosome compared to sex-reversed males with two proto-X
658 chromosomes. To do this, we implemented the IDP-ASE tool at the gene level with house fly
659 genome annotation release 102 (Scott et al. 2014), following the developers' recommended
660 analysis steps (Deonovic et al. 2017). The IDP-ASE software was supplied with raw and aligned
661 reads created by RNA-seq (Son et al. 2019) and Nanopore sequencing, as well as variant calls
662 (SNPs and small indels) in RNA-seq reads created by GATK. We only used the Nanopore reads
663 for phasing haplotypes in the IDP-ASE run, and not for variant calling, because there was less
664 than 10× coverage across the house fly genome (i.e., too low for reliable variant calling).
665
666 The prepared data from each gene was next run in an MCMC sampling simulation to estimate
667 the haplotype within each gene with a Metropolis-Hastings sampler (Bansal et al. 2008). The
668 software estimates the proportion of each estimated haplotype that contributes to the total
669 expression of the gene (ρ) from each iteration using slice sampling (Neal et al. 2003). A value of
670 $\rho=0.5$ indicates equal expression between two alleles, whereas $\rho<0.5$ or $\rho>0.5$ indicates ASE.
671 The MCMC sampling was run with a 1000 iteration burn-in followed by at least 500 iterations
672 where data were recorded. The actual number of iterations was automatically adjusted by the
673 software during the simulation to produce the best simulation output for quantifying ASE within
674 a gene. The IDP-ASE simulation generated a distribution of ρ for each gene across all post-burn-
675 in iterations, and then it calculated the proportion of iterations with $\rho > 0.5$. This proportion was
676 used to estimate the extent of ASE for each gene. For example, if all iterations for a gene have ρ
677 > 0.5, then the proportion is 1 and the gene has strong evidence for ASE of one allele. Similarly,
678 if all iterations for a gene have $\rho < 0.5$, then the proportion is 0 and the gene has strong evidence
679 for ASE of the other allele. In contrast, if half of the iterations have $\rho > 0.5$ and the other half
680 have $\rho < 0.5$, then the proportion is 0.5 and there is not any evidence for ASE. In our subsequent
681 analysis, we only included genes with at least 10 mapped reads combined across three RNA-seq
682 libraries from the genotype under consideration.

683

684 We used the output of IDP-ASE to compare expression of the III^M (proto-Y) and III (proto-X)
685 alleles in genotypic males. IDP-ASE only quantifies ASE within bi-allelic loci, so we only
686 included genes with heterozygous sites within transcripts in genotypic (III^M/III) or sex-reversed
687 (III/III) males. In addition, we removed heterozygous variants with the same genotype in
688 genotypic and sex-reversed males because they do not allow us discriminate between the proto-Y
689 and proto-X alleles. Removing these variants may have also sped up the simulation times, but
690 this was not rigorously investigated. To discriminate between the III^M and III alleles, we used
691 haplotypes estimated during IDP-ASE runs and genotypes inferred from GATK for genotypic
692 (III^M/III) and sex-reversed (III/III) males. For example, using genotypes called using GATK
693 from the RNA-seq data, we first identified sites with heterozygous alleles in genotypic males and
694 homozygous alleles in sex-reversed males. Next, we inferred the allele in common between
695 genotypic and sex-reversed as the III allele, and the other allele that is unique to genotypic males
696 as the III^M allele. Lastly, we matched those sites to the haplotypes estimated by IDP-ASE to
697 quantify ASE within each genotype.
698
699

700 **Acknowledgements and funding information**

701 We thank Leo W. Beukeboom and Daniel Bopp for collaborating on the RNAi knockdown and
702 valuable discussions, and Fei Yuan for help developing a python script used to calculate the
703 percentage of heterozygous variants. This work was supported by a Grant-in-Aid of Research
704 from the National Academy of Sciences, administered by Sigma Xi, The Scientific Research
705 Society (grant number G2018100198487895 to J.H.S. and R.P.M.) and by the National Science
706 Foundation (grant numbers OISE-1444220 and DEB-1845686 to R.P.M.). Analysis of RNA-seq
707 and Oxford Nanopore sequencing data were performed on the Maxwell and Sabine clusters in
708 the Research Computing Data Core at the University of Houston. The Oxford Nanopore long
709 read data used in the study are available from the National Center for Biotechnology Information
710 Sequence Read Archive under BioProject accession PRJNA620357 (BioSample accessions
711 SAMN14518459 for sex-reversed males and SAMN14518460 for genotypic males). The III^M
712 male genome assembly is available from the National Center for Biotechnology Information

713 Whole Genome Shotgun accession JACCFA000000000 under BioProject accession
714 PRJNA620357.

715 # References

716 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
717 J Mol Biol. 215(3):403–410.
718 Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome
719 degeneration. Nat Rev Genet. 14(2):113–124.
720 Bachtrog D, Mahajan S, Bracewell R. 2019. Massive gene amplification on a recently formed
721 Drosophila Y chromosome. Nat Ecol Evol. 3(11):1587–1597.
722 Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J,
723 Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? PLoS Biol.
724 12(7):e1001899.
725 Bansal V, Halpern AL, Axelrod N, Bafna V. 2008. An MCMC algorithm for haplotype assembly
726 from whole-genome sequence data. Genome Res. 18(8):1336–1346.
727 Bergero R, Qiu S, Forrest A, Borthwick H, Charlesworth D. 2013. Expansion of the pseudo-
728 autosomal region and ongoing recombination suppression in the Silene latifolia sex
729 chromosomes. Genetics. 194(3):673–686.
730 Beukeboom LW, Perrin N. 2014. The evolution of sex determination. Oxford University Press,
731 USA.
732 Boyes JW, Corey MJ, Paterson HE. 1964. Somatic chromosomes of higher diptera: IX.
733 Karyotypes of some muscid species. Can J Zool. 42(6):1025–1036.
734 Bull JJ. 1983. Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing
735 Company, Inc.
736 Carvalho AB, Clark AG. 2013. Efficient identification of y chromosome sequences in the human
737 and drosophila genomes. Genome Res. 23(11):1894–1907.
738 Carvalho AB, Koerich LB, Clark AG. 2009. Origin and evolution of Y chromosomes:
739 Drosophila tales. Trends Genet. 25(6):270–277.
740 Carvalho AB, Lazzaro BP, Clark AG. 2000. Y chromosomal fertility factors kl-2 and kl-3 of
741 Drosophila melanogaster encode dynein heavy chain polypeptides. Proc Natl Acad Sci.
742 97(24):13239–13244.
743 Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. Philos Trans R
744 Soc London Ser B Biol Sci. 355(1403):1563–1572.
745 Charlesworth B, Jordan CY, Charlesworth D. 2014. The evolutionary dynamics of sexually
746 antagonistic mutations in pseudoautosomal regions of sex chromosomes. Evolution (N Y).
747 68(5):1339–1350.
748 Charlesworth D. 2017. Evolution of recombination rates between sex chromosomes. Philos
749 Trans R Soc B Biol Sci. 372(1736):20160456.
750 Charlesworth D. 2018. Does sexual dimorphism in plants promote sex chromosome evolution?
751 Environ Exp Bot. 146:5–12.
752 Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex
753 chromosomes. Heredity (Edinb). 95(2):118.
754 Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila
755 melanogaster models of human disease. Nat Genet. 39(6):715–720.
756 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012.
757 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff.
758 Fly (Austin). 6(2):80–92.
759 Darolti I, Wright AE, Sandkam BA, Morris J, Bloch NI, Farré M, Fuller RC, Bourne GR, Larkin
760 DM, Breden F, et al. 2019. Extreme heterogeneity in sex chromosome differentiation and dosage

761 compensation in livebearers. Proc Natl Acad Sci U S A. 116(38):19031–19036.
762 Dean R, Zimmer F, Mank JE. 2014. The potential role of sexual conflict and sexual selection in
763 shaping the genomic distribution of mito-nuclear genes. Genome Biol Evol. 6(5):1096–1104.
764 Deonovic B, Wang Y, Weirather J, Wang X-J, Au KF. 2017. IDP-ASE: Haplotyping and
765 quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing.
766 Nucleic Acids Res. 45(5):e32--e32.
767 Diggle CP, Moore DJ, Mali G, zur Lage P, Ait-Lounis A, Schmidts M, Shoemark A, Garcia
768 Munoz A, Halachev MR, Gautier P, et al. 2014. HEATR2 Plays a Conserved Role in Assembly
769 of the Ciliary Motile Apparatus. PLoS Genet. 10(9):e1004577.
770 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
771 TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15–21.
772 van Doorn GS, Kirkpatrick M. 2007. Turnover of sex chromosomes induced by sexual conflict.
773 Nature. 449(7164):909.
774 van Doorn GS, Kirkpatrick M. 2010. Transitions between male and female heterogamety caused
775 by sex-antagonistic selection. Genetics. 186(2):629–645.
776 Drown DM, Preuss KM, Wade MJ. 2012. Evidence of a paucity of genes that interact with the
777 mitochondrion on the X in mammals. Genome Biol Evol. 4(8):875–880.
778 Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y
779 chromosomes. PLoS Genet. 15(7):e1008251.
780 Feldmeyer B, Pen I, Beukeboom LW. 2010. A microsatellite marker linkage map of the
781 housefly, Musca domestica: Evidence for male recombination. Insect Mol Biol. 19(4):575–581.
782 Gallach M, Chandrasekaran C, Betran E. 2010. Analyses of nuclearly encoded mitochondrial
783 genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic
784 conflict in Drosophila. Genome Biol Evol. 2:835–850.
785 Garcia-Moreno J, Mindell DP. 2000. Rooting a phylogeny with homologous genes on opposite
786 sex chromosomes (gametologs): A case study using avian CHD. Mol Biol Evol. 17(12):1826–
787 1832.
788 Gemmell NJ, Metcalf VJ, Allendorf FW. 2004. Mother's curse: The effect of mtDNA on
789 individual fitness and population viability. Trends Ecol Evol. 19(5):238–244.
790 Gethmann RC. 1988. Crossing over in males of higher diptera (brachycera). J Hered. 79(5):344–
791 350.
792 Ghosh D. 2000. Object-oriented transcription factors database (ooTFD). Nucleic Acids Res.
793 28(1):308–310.
794 Gu L, Walters JR. 2017. Evolution of sex chromosome dosage compensation in animals: A
795 beautiful theory, undermined by facts and bedeviled by details. Genome Biol Evol. 9(9):2461–
796 2476.
797 Hahne F, Ivanek R. 2016. Visualizing genomic data using Gviz and Bioconductor. In: Mathé E,
798 Davis S, editors. Statistical Genomics: Methods and Protocols. New York (NY): Springer. p.
799 335–351.
800 Hamm RL, Meisel RP, Scott JG. 2015. The evolving puzzle of autosomal versus Y-linked male
801 determination in Musca domestica. G3 Genes, Genomes, Genet. 5(3):371–384.
802 Hediger M, Henggeler C, Meier N, Perez R, Saccone G, Bopp D. 2010. Molecular
803 characterization of the key switch F provides a basis for understanding the rapid divergence of
804 the sex-determining pathway in the housefly. Genetics. 184(1):155–170.
805 Hediger M, Minet AD, Niessen M, Schmidt R, Hilfiker-Kleiner D, Çakir Ş, Nöthiger R,
806 Dübendorfer A. 1998. The male-determining activity on the Y chromosome of the housefly

807 (Musca domestica L.) consists of separable elements. Genetics. 150(2):651–661.
808 Hughes JF, Skaletsky H, Pyntikova T, Graves TA, Van Daalen SKM, Minx PJ, Fulton RS,
809 McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human y chromosomes are
810 remarkably divergent in structure and gene content. Nature. 463(7280):536–539.
811 Kaiser VB, Zhou Q, Bachtrog D. 2011. Nonrandom gene loss from the drosophila miranda neo-
812 Y chromosome. Genome Biol Evol. 3:1329–1337.
813 Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. 12(4):656–664.
814 Koerich LB, Wang X, Clark AG, Carvalho AB. 2008. Low conservation of gene content in the
815 Drosophila Y chromosome. Nature. 456(7224):949–951.
816 Krasovec M, Kazama Y, Ishii K, Abe T, Filatov DA. 2019. Immediate Dosage Compensation Is
817 Triggered by the Deletion of Y-Linked Genes in Silene latifolia. Curr Biol. 29(13):2214–2221.
818 Kurek R, Reugels AM, Glaätzer KH, Bünemann H. 1998. The Y chromosomal fertility factor
819 Threads in Drosophila hydei harbors a functional gene encoding an axonemal dynein $β$ heavy
820 chain protein. Genetics. 149(3):1363–1376.
821 Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics.
822 34(18):3094–3100.
823 Mahajan S, Bachtrog D. 2017. Convergent evolution of y chromosome gene content in flies. Nat
824 Commun. 8(1):1–13.
825 Mahajan S, Wei KHC, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young
826 Drosophila Y chromosome using single-molecule sequencing and chromatin conformation
827 capture. PLoS Biol. 16(7):e2006348.
828 Mank JE. 2013. Sex chromosome dosage compensation: Definitely not for everyone. Trends
829 Genet. 29(12):677–683.
830 Mank JE. 2017. The transcriptional architecture of phenotypic dimorphism. Nat Ecol Evol.
831 1(1):1–7.
832 Mcdonald I c, Evenson P, Nickel CA, Johnson OA. 1978. House fly genetics: isolation of a
833 female determining factor on chromosome 4. Ann Entomol Soc Am. 71(5):692–694.
834 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
835 Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce
836 framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
837 Meisel RP, Davey T, Son JH, Gerry AC, Shono T, Scott JG. 2016. Is Multifactorial Sex
838 Determination in the House Fly, Musca domestica L., Stable Over Time? J Hered. 107(7):615–
839 625. doi:10.1093/jhered/esw051.
840 Meisel RP, Gonzales CA, Luu H. 2017. The house fly Y Chromosome is young and minimally
841 differentiated from its ancient X Chromosome partner. Genome Res. 27(8):1417–1426.
842 Meisel RP, Olafson PU, Adhikari K, Guerrero FD, Konganti K, Benoit JB. 2020. Sex
843 Chromosome Evolution in Muscid Flies. G3 (Bethesda). 10(4):1341–1352.
844 Meisel RP, Scott JG. 2018. Using genomic data to study insecticide resistance in the house fly,
845 Musca domestica. Pestic Biochem Physiol. 151:76–81.
846 Montgomery SH, Mank JE. 2016. Inferring regulatory change from gene expression: the
847 confounding effects of tissue scaling. Mol Ecol. 25(20):5114–5128.
848 Neal RM. 2003. Slice sampling. Ann Stat. 31(3):705–767.
849 Orzack SH, Sohn JJ, Kallman KD, Levin SA, Johnston R. 1980. Maintenance of the three sex
850 chromosome polymorphism in the platyfish, Xiphophorus maculatus. Evolution (N Y).
851 34(4):663–672.
852 Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene

853 expression. Nat Rev Genet. 14(2):83–87.
854 Picard CJ, Johnston JS, Tarone AM. 2012. Genome Sizes of Forensically Relevant Diptera. J
855 Med Entomol. 49(1):192–197.
856 Ponnikas S, Sigeman H, Abbott JK, Hansson B. 2018. Why Do Sex Chromosomes Stop
857 Recombining? Trends Genet. 34(7):492–503.
858 Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R,
859 Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance.
860 Nature. 530(7589):228–232.
861 Rand DM, Clark AG, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in
862 Drosophila melanogaster. Genetics. 159(1):173–187.
863 Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution (N Y).
864 38(4):735–742.
865 Rice WR. 1987. The accumulation of sexually antagonistic genes as a selective agent promoting
866 the evolution of reduced recombination between primitive sex chromosomes. Evolution (N Y).
867 41(4):911–914.
868 Rice WR. 1996a. Evolution of the Y sex chromosome in animals. Bioscience. 46(5):331–343.
869 Rice WR. 1996b. Sexually antagonistic male adaptation triggered by experimental arrest of
870 female evolution. Nature. 381(6579):232–234.
871 Roberts RB, Ser JR, Kocher TD. 2009. Sexual conflict resolved by invasion of a novel sex
872 determiner in Lake Malawi cichlid fishes. Science (80-). 326(5955):998–1001.
873 Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods.
874 17(2):155–158.
875 Sardell JM, Cheng C, Dagilis AJ, Ishikawa A, Kitano J, Peichel CL, Kirkpatrick M. 2018. Sex
876 differences in recombination in sticklebacks. G3 (Bethesda). 8(6):1971–1983.
877 Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK,
878 Kasai S, Leichter CA, et al. 2014. Genome of the house fly, Musca domestica L., a global vector
879 of diseases with adaptations to a septic environment. Genome Biol. 15(10):466.
880 Sharma A, Heinze SD, Wu Y, Kohlbrenner T, Morilla I, Brunner C, Wimmer EA, van de Zande
881 L, Robinson MD, Beukeboom LW, et al. 2017. Male sex in houseflies is determined by Mdmd, a
882 paralog of the generic splice factor gene CWC22. Science (80-). 356(6338):642–645.
883 Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S,
884 Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is
885 a mosaic of discrete sequence classes. Nature. 423(6942):825–837.
886 Soh YQS, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C,
887 Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse y chromosome reveals convergent
888 gene acquisition and amplification on both sex chromosomes. Cell. 159(4):800–813.
889 Son JH, Kohlbrenner T, Heinze S, Beukeboom LW, Bopp D, Meisel RP. 2019. Minimal effects
890 of proto-Y chromosomes on house fly gene expression in spite of evidence that selection
891 maintains stable polygenic sex determination. Genetics. 213(1):313–327.
892 Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific
893 expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics.
894 14(1):536.
895 Vicoso B. 2019. Molecular and evolutionary dynamics of animal sex-chromosome turnover. Nat
896 Ecol Evol. 3:1632–1641.
897 Wei KHC, Bachtrog D. 2019. Ancestral male recombination in Drosophila albomicans produced
898 geographically restricted neo-Y chromosome haplotypes varying in age and onset of decay.

899 PLoS Genet. 15(11):e1008502.
900 Zhou Q, Bachtrog D. 2012a. Chromosome-wide gene silencing initiates y degeneration in
901 drosophila. Curr Biol. 22(6):522–525.
902 Zhou Q, Bachtrog D. 2012b. Sex-specific adaptation drives early sex chromosome evolution in
903 Drosophila. Science (80-). 337(6092):341–345.
904 Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MTP, Zhang G. 2014.
905 Complex evolutionary trajectories of sex chromosomes across bird taxa. Science (80-).
906 346(6215):1246338.
907 Zimmer F, Harrison PW, Dessimoz C, Mank JE. 2016. Compensation of dosage-sensitive genes
908 on the chicken Z chromosome. Genome Biol Evol. 8(4):1233–1242.
909

Figure legends

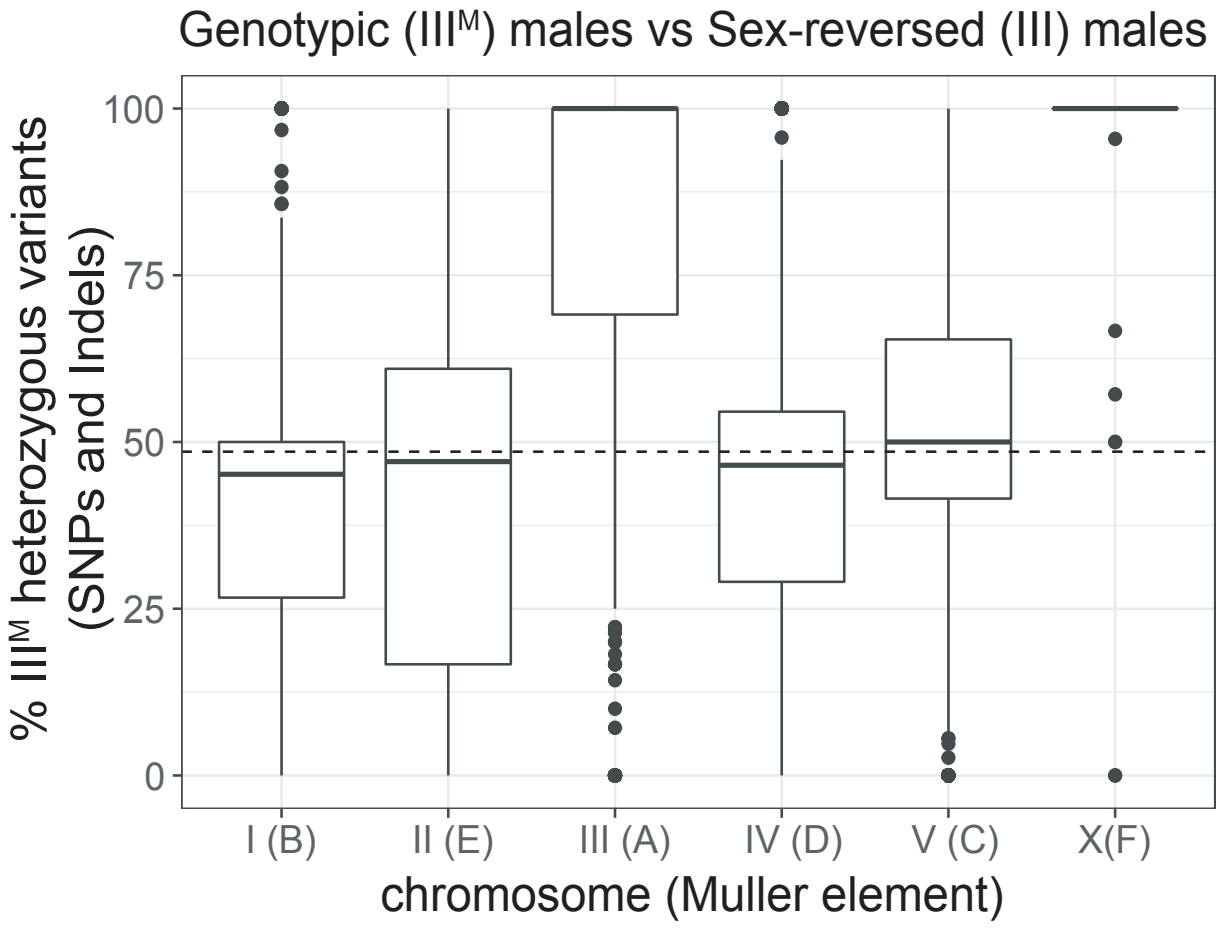
Figure 1. Elevated heterozygosity on the third and X chromosomes in genotypic (III^M/III) males relative to sex-reversed (III/III) males. The boxplots show the distributions of the percent of heterozygous variants per gene in the genotypic males relative to the sex-reversed males (% III^M heterozygous variants) on each chromosome. Muller element nomenclature for each chromosome is shown in parentheses (Meisel and Scott 2018). See Materials and Methods for the calculation of % III^M heterozygous variants. Values more than 50% indicate more heterozygous variants in genotypic (III^M/III) males, and less than 50% indicates more heterozygous variants in sex-reversed males. The median across autosomes I, II, IV, and V is represented by a dashed line.

Figure 2. Two proto-X and proto-Y loci identified in the III^M genome assembly. (A) One locus was identified with a truncated *Mdmd* on the proto-Y (ctg2382) and the same three genes on both the proto-X (ctg1607) and proto-Y. (B) One locus was identified with four genes on both the proto-X (ctg1519) and proto-Y (ctg2522). The two contigs (ctg2522 and ctg1519) were assigned to the proto-Y and proto-X based on sequences that are enriched in the male relative to female reads (see Supplementary Figure 1).

Figure 3. Evidence for moderately elevated ASE on the third (proto-Y) chromosome in III^M males. (A) Proportions of genes with ASE in genotypic (III^M) or sex-reversed (III) males on each chromosome. There is not a significant difference on any chromosome between the two genotypes. (B) Proportions of genes with ASE in genotypic males and non-ASE in sex-reversed males on the third chromosome and all other chromosomes (left two bars). Proportions of genes with non-ASE in genotypic males and ASE in sex-reversed males on the third chromosome and all other chromosomes (right two bars). The asterisk indicates a significant difference ($p<0.05$) in the number of genes in these categories as determined by Fisher's exact test.

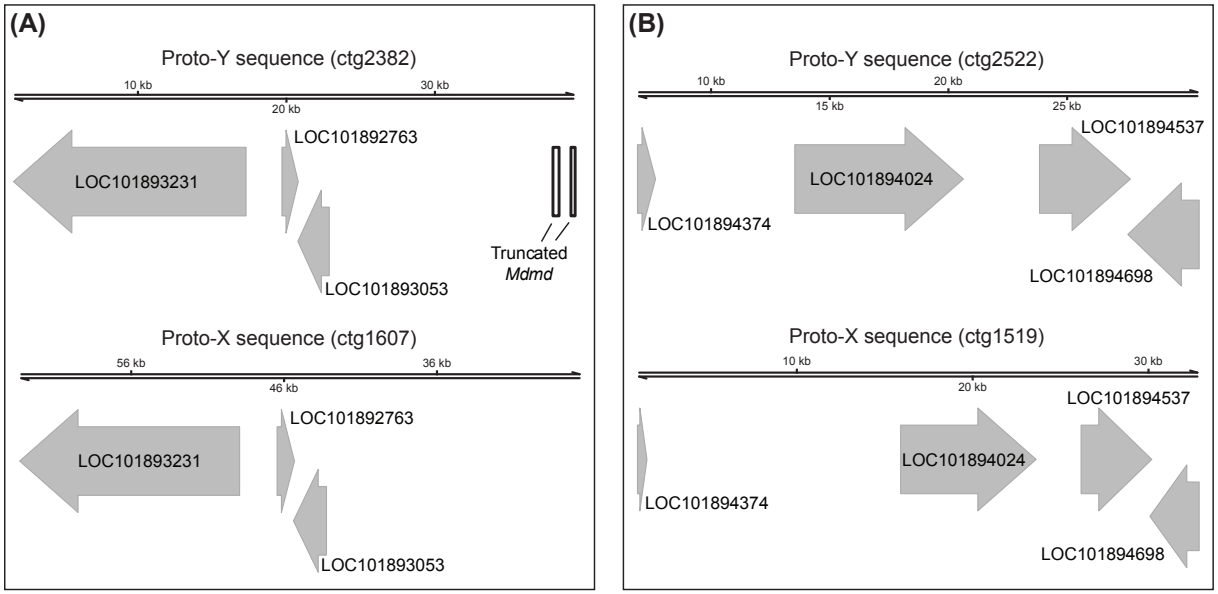
Figure 4. Allele-specific expression (ASE) of *Md-HEATR2*. (A) Diagnostic variable sites for ASE in the *Md-HEATR2* gene are based on haplotypes estimated in IDP-ASE. Read depth is measured as fragments per million mapped reads (FPM) in III^M males, sex-reversed (SR) males, and the III^M or III allele in III^M males. (B) Variable sites that differ between genotypic (III^M) males and sex-reversed males (triangles) across 1,273 base pairs upstream of *Md-HEATR2* were identified using Oxford Nanopore long reads only. (C) Transcription factor (TF) binding motifs predicted within 1,273 base pairs upstream of *Md-HEATR2*. The starting position of each motif is given, along with its length (in parentheses). None of the variable sites overlap with predicted TF motifs. All positions are coordinates in scaffold NW_004764965.1 of the reference genome assembly (Scott et al. 2014).

949 **Figure 1**



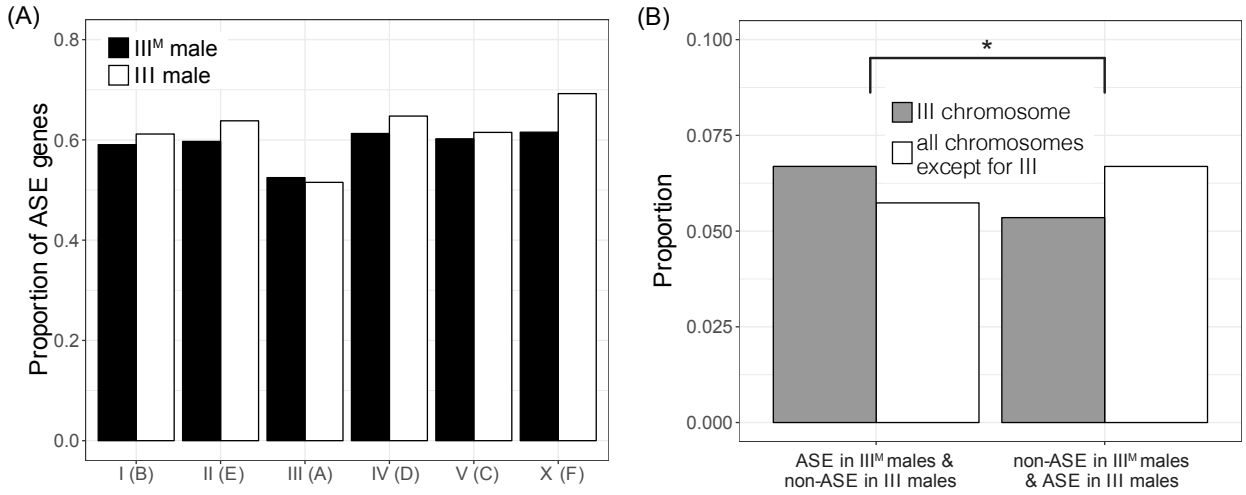
950

951 **Figure 2**



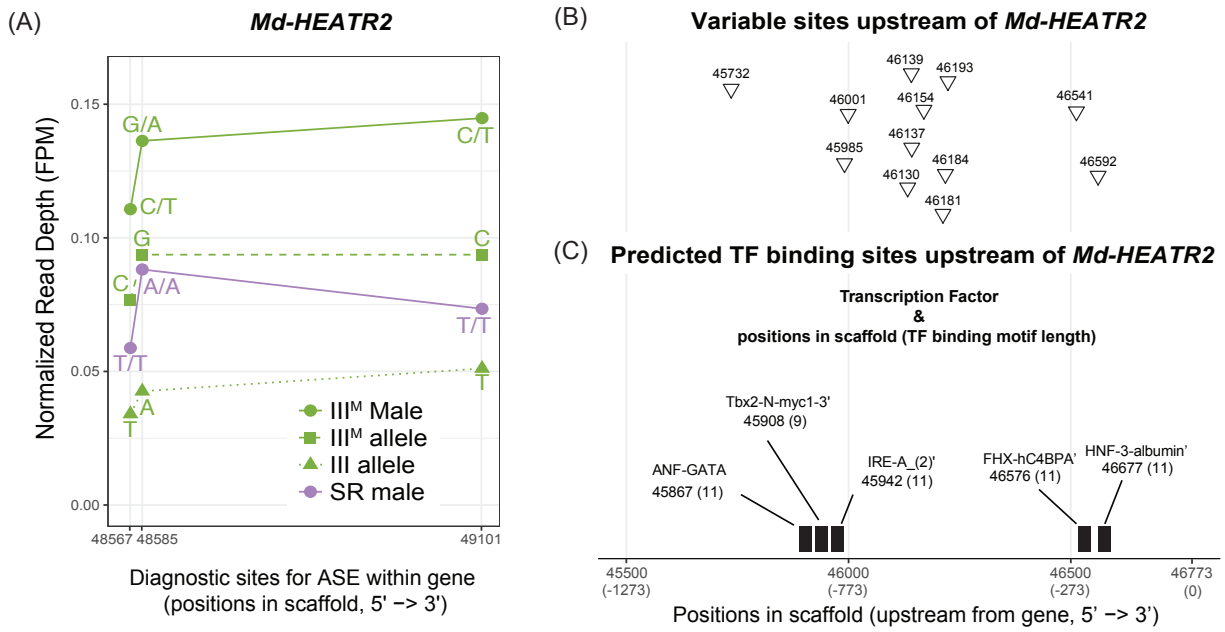
952

953 **Figure 3**



954

955 **Figure 4**



956

Supplementary Materials

Identification of two candidate proto-X/proto-Y loci

We identified one contig containing *Mdmd* in the III^M assembly (ctg2382) that contains the same three genes as another contig (ctg1607) without *Mdmd* (Figure 2A). We infer the contig with *Mdmd* to be the proto-Y sequence, and the one without to be from the proto-X. There are two additional contigs containing *Mdmd* in our III^M genome assembly, but we did not find any corresponding proto-X contigs for these (Supplementary Figure 1). These other two contigs with *Mdmd* either contain loci without sufficient X-Y divergence to assemble into separate proto-X and proto-Y contigs, or we did not sequence deep enough to capture enough proto-X reads to assemble the proto-X gametolog. The male-determining regions of the house fly proto-Y chromosomes contain both complete and truncated copies of *Mdmd* (Sharma et al. 2017), and ctg2382 contains two truncated copies of *Mdmd*. The genomic region corresponding to this contig is found in a single scaffold of the reference genome (NW_004774683.1), which was previously assigned to the third chromosome (Meisel and Scott 2018). There are three genes at this locus, and they are present in the same order in the reference genome, the proto-Y contig, and proto-X contig (Figure 2A).

We found the second locus using a *k*-mer comparison approach to test for contigs containing sequences that are enriched in males relative to females (Carvalho and Clark 2013). Here, we compared *k*-mers in our Nanopore III^M male reads and the Illumina reads used to generate the reference (female) genome assembly (Scott et al. 2014) to our III^M male assembly. We then calculated the percent of each contig in our III^M assembly that is unmatched by female reads (%UFR), which we expect to be 100% for a Y chromosome contig that only contains male-specific *k*-mers. We did not identify any contigs with greater than 27.4%UFR (Supplementary Figure 1), which is consistent with the minimal sequence divergence between the proto-Y and proto-X chromosomes (Meisel et al. 2017). To determine an expectation for what constitutes a %UFR of a proto-Y contig in our assembly, we calculated %UFR for the three contigs that each contain copies of *Mdmd* (Supplementary Figure 1). The largest %UFR of any contig with

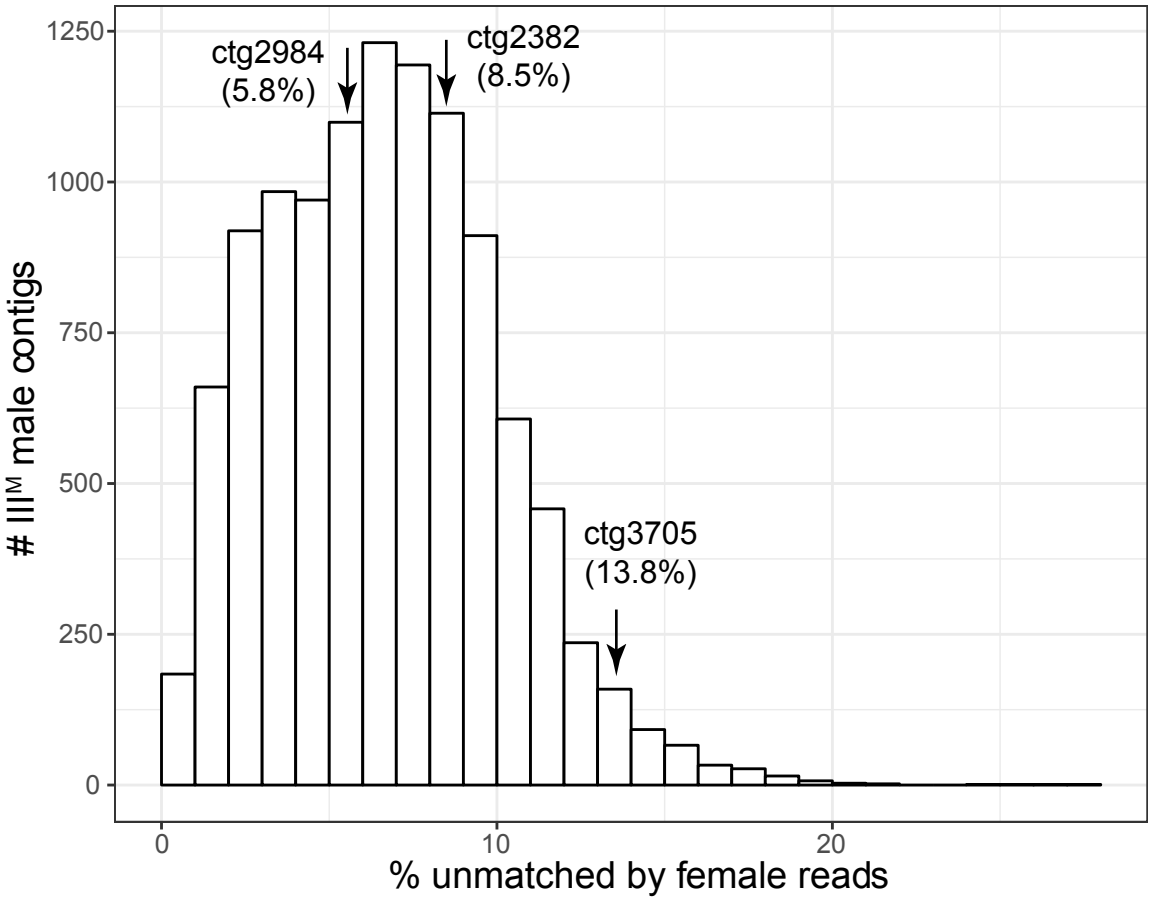
Mdmd is 13.8% (ctg3705). We used this value as a threshold to find proto-Y contigs that do not contain *Mdmd*.

We found 275 contigs with a %UFR>13.8, nine of which (Supplementary Table 1) contain protein-coding genes that were previously assigned to the third chromosome (Meisel and Scott 2018). One gene (LOC101898200, which encodes histone H3) was found on six different contigs with %UFR>13.8 (and an additional 27 contigs with %UFR<13.8). Histone gene clusters are a common feature of most metazoan genomes, and the multiple copies are under tight regulatory control (McKay et al. 2015; Duronio and Marzluff 2017).These contigs with high %UFR and copies of histone H3 genes are either false positives or not of substantial phenotypic relevance. Two of the other contigs with %UFR>13.8 (ctg3539 and ctg8407) each contain 1-2 genes that are found as multiple copies on other contigs (all with a %UFR<13.8). One of those genes (LOC109612838 on ctg3539) encodes a protein with predicted reverse transcriptase and RNase H activity, suggesting that it is a retrotransposon (Finnegan 2012). Another gene (LOC109613819 on ctg8407) encodes a protein with zinc finger and integrase domains, also consistent with a transposable element (Volff 2006). We exclude these contigs from our subsequent analysis because they likely reflect transposable element expansions within the house fly genome. We only found one contig with a %UFR>13.8 (ctg2522, which has 16.9%UFR) that has a single corresponding contig (ctg1519) with the same four genes (Supplementary Table 1). We assigned ctg2522 to the proto-Y chromosome and ctg1519 to the proto-X (Figure 2B). All four genes shared by both contigs are mapped to the third chromosome in the reference genome. These four genes are present in the same order and same orientation in the reference genome (on scaffold NW_004764689.1), the proto-Y contig, and proto-X contig.

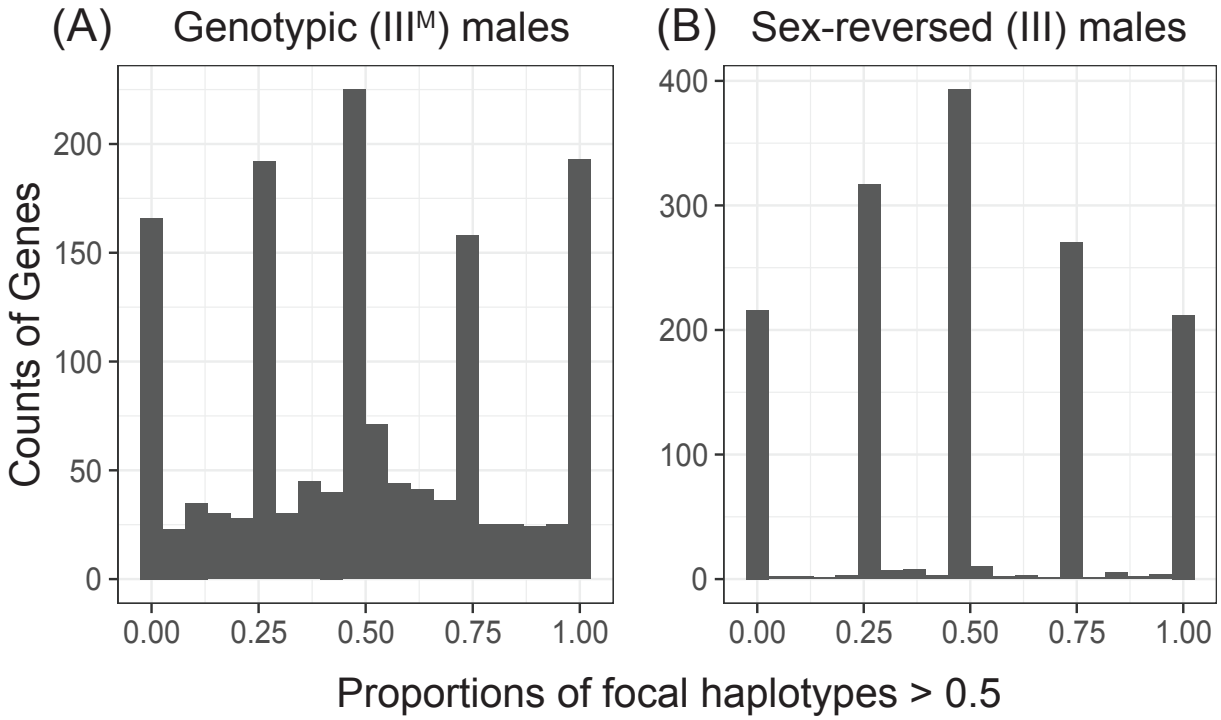
DNA isolation with phenol/chloroform protocol

A single genotypic male and a single sex-reversed male with detached wings were each transferred to a 1.5 mL tube with 0.5 mL homogenization buffer (4.1 g sucrose, 15 mL 1M Tris-HCl pH 8.0, 0.5M EDTA, 100 mL dH_2O) and then homogenized using a pestle set into a tissue grinder homogenizer. To each tube we added 40 uL of 10% SDS and 2.5 uL of 10 mg/mL Proteinase K, and then we incubated the tube at 65°C for 30 min. We next added 2 uL of 4

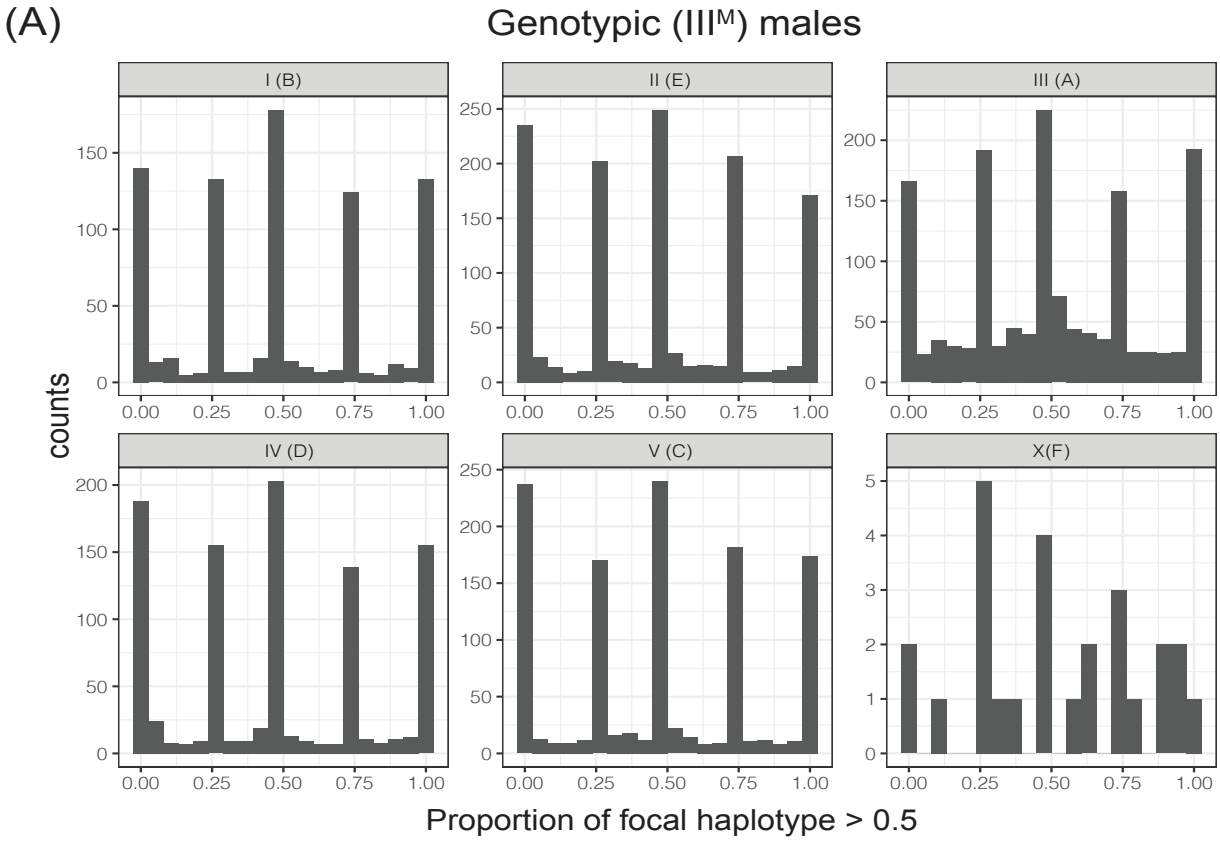
mg/mL RNase to each tube and incubated at 37°C for 15 min. We added 48 uL of 5M KAc to each tube and placed on ice for 30 min. Then we centrifugated the tubes at 14000 rpm for 10 min at 4°C, and the supernatant was transferred into a new tube using a wide-bore pipette tip (all subsequent steps of transfer and mixing during DNA extraction were also done with wide-bore pipette tips to prevent DNA shearing). We added 250 uL phenol and 250 uL chloroform to the extracted supernatant in the new tube, mixed briefly, spun at 14000 rpm for 15 min at 4°C, and then transferred the supernatant into a new tube. We next added 500 uL chloroform to the supernatant in the new tube, mixed well, and spun a 14000 rpm for 5 min at 4°C. We then transferred the supernatant into a new tube. We added 40 uL of 3M NaAc and 800 uL of 95% ethanol to the supernatant in the new tube, mixed briefly, spun at 14000 rpm for 15 min at 4°C, and then carefully poured off all supernatant. We next added 800 uL of 70% ethanol to the remaining pellet, mixed briefly to wash the pellet, spun at 14000 rpm for 15 min at 4°C, removed the supernatant, and then resuspended the pellet in 30 uL of nuclease-free water.

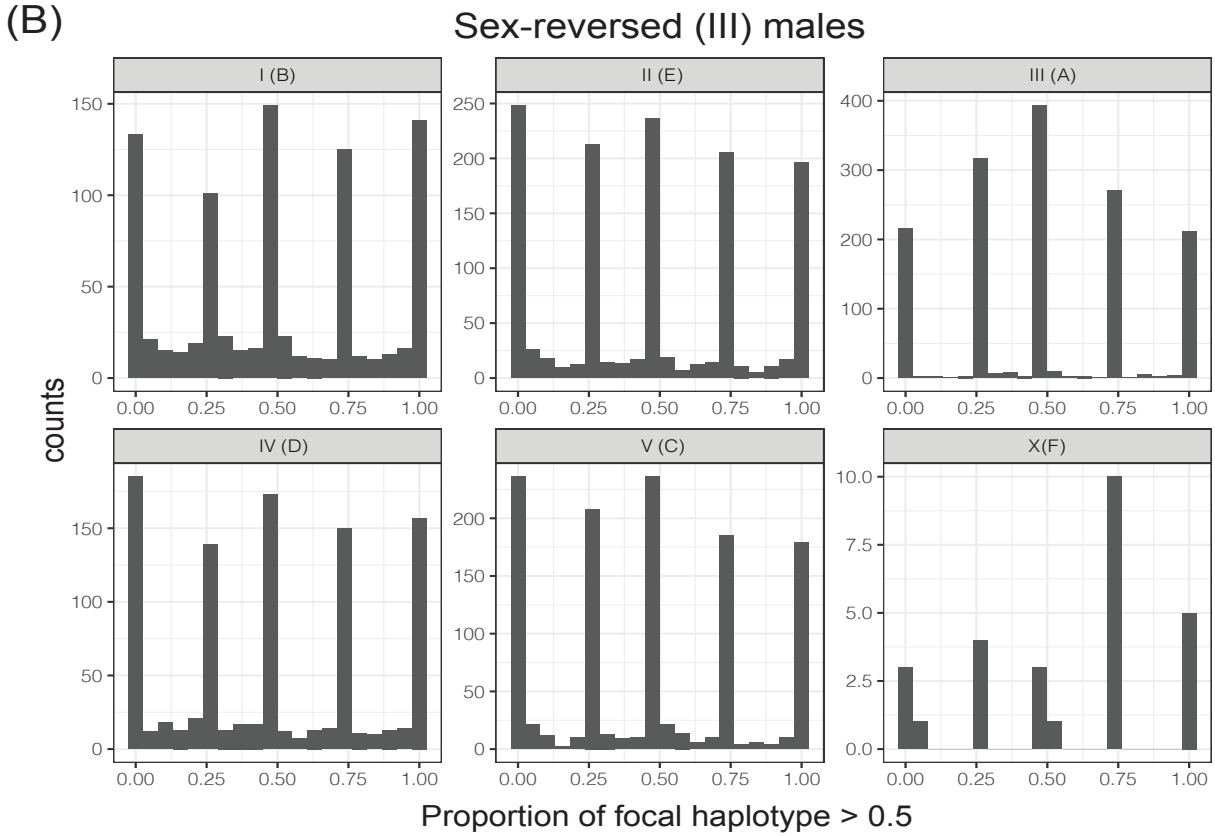


Supplementary Figure 1. Histograms of the percent unmatched by female reads (%UFR) for all contigs in the III^M male genome assembly. Three contigs with *Mdmd* are shown, with their %UFR in parentheses.



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. k-mer comparison in the III^M male genome assembly
Supplementary Data 4. IDP-ASE (allele-specific expression) output for the genotypic males
Supplementary Data 5. IDP-ASE (allele-specific expression) output for the sex-reversed males

Supplementary Data 1 and 2 are available at Dryad (https://doi.org/10.5061/dryad.280gb5mnk or https://datadryad.org/stash/share/6f4xn0dRAu8j2SQXCmqKICv95gYfpGzlhoR5SYLJZy0)

References

Carvalho AB, Clark AG. 2013. Efficient identification of y chromosome sequences in the human and drosophila genomes. Genome Res. 23(11):1894–1907.
Duronio RJ, Marzluff WF. 2017. Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol. 14(6):726–738.
Finnegan DJ. 2012. Retrotransposons. Curr Biol. 22(11):R432–R437.
McKay DJ, Klusza S, Penke TJR, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW, Tatomer DC, Lieb JD, et al. 2015. Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell. 32(3):373–386.
Meisel RP, Scott JG. 2018. Using genomic data to study insecticide resistance in the house fly, Musca domestica. Pestic Biochem Physiol. 151:76–81.
Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA, et al. 2014. Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. Genome Biol. 15(10):466.
Sharma A, Heinze SD, Wu Y, Kohlbrenner T, Morilla I, Brunner C, Wimmer EA, van de Zande L, Robinson MD, Beukeboom LW, et al. 2017. Male sex in houseflies is determined by Mdmd, a paralog of the generic splice factor gene CWC22. Science (80-). 356(6338):642–645.
Volff JN. 2006. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays. 28(9):913–922.