Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*

Josh Hough[1], Jesse D. Hollister, Wei Wang, Spencer C. H. Barrett, and Stephen I. Wright

Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada M5S 3B2

Edited by James A. Birchler, University of Missouri-Columbia, Columbia, MO, and approved April 18, 2014 (received for review October 11, 2013)

Heteromorphic sex chromosomes have originated independently in many species, and a common feature of their evolution is the degeneration of the Y chromosome, characterized by a loss of gene content and function. Despite being of broad significance to our understanding of sex chromosome evolution, the genetic changes that occur during the early stages of Y-chromosome degeneration are poorly understood, especially in plants. Here, we investigate sex chromosome evolution in the dioecious plant *Rumex hastatulus*, in which X and Y chromosomes have evolved relatively recently and occur in two distinct systems: an ancestral XX/XY system and a derived XX/XY_1Y_2 system. This polymorphism provides a unique opportunity to investigate the effect of sex chromosome age on patterns of divergence and gene degeneration within a species. Despite recent suppression of recombination and low X-Y divergence in both systems, we find evidence that Y-linked genes have started to undergo gene loss, causing ∼28% and ∼8% hemizygosity of the ancestral and derived X chromosomes, respectively. Furthermore, genes remaining on Y chromosomes have accumulated more amino acid replacements, contain more unpreferred changes in codon use, and exhibit significantly reduced gene expression compared with their X-linked alleles, with the magnitude of these effects being greatest for older sex-linked genes. Our results provide evidence for reduced selection efficiency and ongoing Y-chromosome degeneration in a flowering plant, and indicate that Y degeneration can occur soon after recombination suppression between sex chromosomes.

molecular evolution | sex linkage | dioecy

Systems of sex determination involving X and Y chromosomes have evolved multiple times in both plants and animals, with Y chromosomes having lost much of their genetic function in many species (1–3). Evidence of DNA sequence homology between X- and Y-linked gene pairs in flowering plants (4–7) and fish (8) supports the idea that sex chromosomes have evolved from autosomes and subsequently diverged following the suppression of recombination between genes involved in sex determination. Evolutionary models predict that when regions of suppressed recombination evolve on Y chromosomes, the associated reduction in the effectiveness of selection should lead to a pattern of Y-chromosome degeneration in which genes carried on the Y become impaired in function and are eventually lost (1–3). The well-studied Y chromosomes in humans and *Drosophila melanogaster*, for example, show clear signs of degeneration: They almost completely lack homology to the X chromosome, exhibit a highly heterochromatic chromatin structure consisting largely of repetitive and ampliconic DNA, and carry few remaining protein-coding genes (9–13).

Recent genomic studies of sex chromosomes in humans, rhesus macaques, and chimpanzees (12, 13) have provided detailed information regarding the genetic structure and gene content of Y chromosomes, shedding light on the processes contributing to their deterioration. However, we still know little about the changes characterizing the early stages of Y-chromosome degeneration or the time scales over which they occur. This situation arises because sex chromosomes in these well-studied mammalian species evolved >200 Mya (14, 15), and therefore provide few clues about their early evolutionary history. Genomic studies of younger plant Y chromosomes (16–19) and *Drosophila* neo-Y chromosomes (20–23), where degeneration is in progress, thus provide excellent opportunities to gain insight into the early processes involved in sex chromosome divergence.

Here, we investigate X- and Y-chromosome evolution in the annual, dioecious plant *Rumex hastatulus* (Polygonaceae). Sex chromosomes in *R. hastatulus* represent an interesting case of the recent evolution of sex chromosome heteromorphism, with age estimates based on nuclear and chloroplast phylogenies suggesting that sex chromosomes evolved within the past 15–16 million years (24). The presence of a neo-Y sex chromosome system (XX/XY_1Y_2), recently derived from an XX/XY system following a fusion of the X chromosome and a former autosome (25), provides a unique opportunity to contrast patterns of sex chromosome evolution between different sex chromosome systems and to investigate the effect of sex chromosome age on patterns of divergence and degeneration within a species. We used high-throughput transcriptome sequencing of multiple parent–offspring families and an analysis of SNP segregation patterns to identify and compare the expression and molecular evolution of sex-linked genes, with the aim of determining whether Y-linked genes are accumulating deleterious mutations, exhibit reduced expression, or have undergone gene loss.

Results and Discussion

We identified genes linked to sex chromosomes by tracing the inheritance of SNPs from parents to first generation (F_1) progeny in two crosses, one from each sex chromosome system (XX/XY and XX/XY_1Y_2). We identified genes in which SNPs

Significance

Evolutionary theory predicts that in dioecious organisms with sex chromosomes, suppressed X-Y recombination should lead to a loss of Y-chromosome gene content and function. However, the extent to which this process occurs in plants, where sex chromosomes evolved relatively recently, is poorly understood. We tested for Y degeneration in *Rumex hastatulus*, an annual plant that has both XY and XY_1Y_2 sex chromosome systems. We found that Y-linked genes are undergoing degeneration despite their recent origin; they show a faster accumulation of amino acid substitutions, contain more unpreferred changes in codon usage, and are reduced in expression relative to X-linked alleles. Significantly, the magnitude of these effects depended on sex chromosome age, being greater for genes that have been nonrecombining for longer.

Author contributions: J.H., S.C.H.B., and S.I.W. designed research; J.H. and S.I.W. performed research; J.H., J.D.H., and W.W. contributed new reagents/analytic tools; J.H., J.D.H., W.W., and S.I.W. analyzed data; and J.H., J.D.H., S.C.H.B., and S.I.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in GenBank's Short Read Archive (SRA) [accession nos. SRP041588 (*Rumex hastatulus*) and SRP041613 (*Rumex bucephalophorus*)].

[1]To whom correspondence should be addressed. E-mail: josh.hough@utoronto.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319227111/-/DCSupplemental.

segregated in a manner characteristic of sex linkage, with Y alleles transmitted from fathers to sons and X alleles transmitted from fathers to daughters, a method validated in previous studies (16, 17). This approach allowed us to identify 698 genes with four or more sex-linked SNPs in XX/XY populations and 1,298 such genes in XX/XY_1Y_2 populations (Table 1 and *SI Appendix*, Table S1). Approximately 70% of sex-linked genes from the XY system were identified in the XY_1Y_2 system, and ~40% of genes in the XY_1Y_2 system were shared with the XY system. This suggests that the XY_1Y_2 system has acquired many new sex-linked genes since the fusion event, and our analysis allowed us to identify a set of 488 "old" sex-linked genes shared between the systems, as well as 607 "young" genes unique to the XY_1Y_2 system.

Cytological measurements of X-chromosome size in *R. hastatulus* suggest that the X is ~20% of the diploid female genome for the XY system and ~30% of the genome in the XY_1Y_2 system (25). Using the estimated number of genes reported in other dicotyledonous plants [28,000 in *Arabidopsis thaliana* (26)], we obtained a rough estimate of the expected number of sex-linked genes of 5,600 and 8,400 for the XY and XY_1Y_2 systems, respectively. Our screen for sex-linked genes using segregating polymorphisms in expressed genes therefore captures ~13% and ~15% of the total number of sex-linked genes for the XY and XY_1Y_2 systems, respectively.

Because some of our candidate sex-linked genes may be in a pseudoautosomal region, and therefore partially recombining with the sex-determining region, we independently sequenced transcriptomes from a single male and female from each of six populations per sex chromosome system and checked for the presence of fixed differences between males and females (Table 1 and *SI Appendix*, Tables S2 and S3). This approach led to validation of ~80% of the sex-linked genes from the XY system, 90% from the XY_1Y_2 system shared with the XY system, but only 28% of the young XY_1Y_2 genes. This suggests that fewer variants have fixed between the neo-sex chromosomes, potentially due to ongoing recombination in a pseudoautosomal region or very recent recombination suppression, with residual shared polymorphism between the chromosomes. For subsequent analyses, we excluded genes without fixed differences between X and Y, as well as a small number of genes with one or more SNPs displaying autosomal segregation (*SI Appendix*, Table S4).

Phylogenetic Relationships and Evolutionary Divergence of Sex-Linked Genes. To investigate relatedness and levels of divergence of sex-linked genes, we obtained additional transcriptome data and identified orthologous sequences from the closest known nondioecious outgroup that lacks sex chromosomes, *Rumex bucephalophorus* (24). We developed a maximum likelihood method to infer the phased X and Y sequences from both sex chromosome systems for each gene. We confirmed the reliability of our method using simulations (*SI Appendix*, Figs. S1 and S2) and constructed phylogenetic trees of these sequences, including the outgroup (*Methods* and *SI Appendix*). Of 354 old sex-linked genes, 150 (42%) X alleles were monophyletic from the two sex chromosome systems, whereas 179 (51%) Y alleles were monophyletic (Fisher's exact test, $P < 0.04$), consistent with the origins of these Y-linked genes predating the divergence of the two sex chromosome systems. Overall, only 78 (22%) exhibited complete reciprocal monophyly for both X and Y between the systems, highlighting their very recent divergence and indicating that a significant proportion of even the old genes may have experienced recent suppression of recombination and some may be pseudoautosomal. Consistent with this, maximum likelihood estimates of synonymous substitution rates (K_s) for both young and old X- and Y-linked genes (Fig. 1) suggest that the majority have low levels of nucleotide divergence, implying that many genes are in an early stage of divergence or experience ongoing recombination.

It is of interest to infer the extent to which sex-linked genes fall into distinct evolutionary strata, which has been found in animal and plant sex chromosomes (e.g., refs. 14, 27, 28) and is characterized by a stratified increase in divergence of X/Y genes with increasing distance from the pseudoautosomal region. We found a range of K_s values for sex-linked genes within each system, which may reflect that recombination suppression occurred at different times for different genes (which is thought to be the underlying cause of strata). In addition, we found significant differences in average branch-specific K_s values when comparing old vs. young X-linked genes (0.00870 and 0.00276, respectively; $P << 10^{-10}$) and old vs. young Y-linked genes (0.0120 and 0.00297, respectively; $P << 10^{-10}$), with the younger sets showing more left-shifted K_s distributions and much lower average K_s values (Fig. 1). Overall, these results highlight that there has been little sequence divergence for young sex-linked genes (the youngest evolutionary stratum), whereas older genes likely include genes that have experienced recent restricted recombination either before or following the divergence between sex chromosome systems, some genes that may still be pseudoautosomal, and genes that have been nonrecombining for a much longer period (i.e., belong to an older evolutionary stratum).

Y Chromosome Gene Loss and Loss of Expression. The relatively recent evolution of recombination suppression and low sequence divergence between many genes on *R. hastatulus* sex chromosomes raises the question of whether Y-linked genes have been lost, or have lost expression relative to X-linked genes. Gene loss has occurred extensively on human and *Drosophila* Y chromosomes (reviewed in ref. 3), and it might be driven by adaptive silencing of Y-linked genes to mask their deleterious effects (22, 29) or, more passively, as a consequence of harmful mutations occurring in regions essential for gene function (30, 31). We inferred the amount of gene loss in *R. hastatulus* by quantifying the percentage of X-linked genes in which SNP segregation patterns indicated hemizygosity in males (Table 1 and *SI Appendix*). Estimates of hemizygosity based only on mRNA sequence data will include genes that have been lost, genes with nonfunctional (nonexpressed) Y-linked copies, and genes that have moved from autosomes to the X chromosome but do not

Table 1. Numbers of identified sex-linked genes in *R. hastatulus*

Gene set	Sex-linked genes with Y-linked copies*	Hemizygous genes	Hemizygous genes,† %
XY system	698 (565)	119	24
XY_1Y_2 shared‡	510 (460)	100	28
XY_1Y_2 unique§	788 (223)	44	8

*Numbers indicate genes with at least four supporting SNPs showing sex-linked segregation and having no SNPs with autosomal segregation. Values in parentheses identify the numbers of genes with at least one fixed X-Y difference in the population sample.

†Estimates of percentage of hemizygous genes were calculated by comparing the number of hemizygous genes and the number of X/Y genes that had at least four segregating X polymorphisms.

‡Shared genes represent genes in the XX/XY_1Y_2 system that were also identified in the XX/XY system.

§Unique genes represent genes identified as unique to the XX/XY_1Y_2 system.

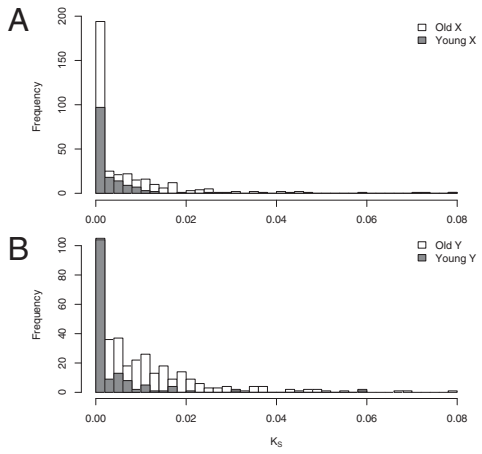

Fig. 1. Synonymous site divergence in sex-linked genes of the XY_1Y_2 system of *R. hastatulus*. Maximum likelihood estimates of lineage-specific rates of per-site synonymous substitution are shown for the X chromosome (*A*) and Y chromosome (*B*). Old sex-linked genes refer to genes that are shared between the ancestral XY system and the derived XY_1Y_2 system. Young sex-linked genes refer to those that are unique to the derived XY_1Y_2 system.

have homologous copies on the Y chromosome. We note that hemizygosity could conceivably be incorrectly inferred using our RNA sequencing (RNAseq)-based approach in cases where X-linked genes have Y-linked copies but are expressed too low to be detected. Such genes would indicate partial Y degeneration rather than genuine gene loss.

By comparing the number of hemizygous genes with the number of X/Y genes with equivalent segregating X-linked polymorphisms (*SI Appendix*), we estimate that the percentage of genes lost from the *R. hastatulus* Y chromosome is as high as 28% (Table 1 and *SI Appendix*, Table S5). We also found that estimates of hemizygosity in XY_1Y_2 males were much lower (8%) than in XY males (Table 1), which is expected because the XX/XY_1Y_2 sex chromosome system has acquired additional X/Y gene pairs, with little time for gene degeneration and loss. Our estimates of the percentage of hemizygosity, although low in comparison to mammalian sex chromosomes [where ~97% of the X chromosome is hemizygous in males (15, 32)], are somewhat higher than other estimates from plants [~20% in *Silene latifolia* (16, 17)] and suggest that Y chromosomes in *R. hastatulus* have already undergone gene loss, despite their relatively recent origin.

We tested for a reduction in expression of young and old Y-linked genes by comparing the ratio of Y/X gene expression in males. Expression was estimated by counting the number of mRNA transcript reads mapping to X/Y SNPs in contigs with four or more such SNPs segregating in F_1 offspring. Because Y-linked alleles in our segregation analysis are identified as alternate alleles at heterozygous sites (with the X allele as the reference), it is important to evaluate the extent of the reduction in the Y/X expression ratio by comparing it with the expression ratio of alternate-to-reference alleles at heterozygous sites throughout the genome. This is necessary because there is an inherent bias toward mapping more reference than alternate alleles (33), and not controlling for this would generate a false signal of lower Y expression or exaggerate signals of truly reduced expression. We therefore tested for reductions in Y/X expression ratios by using the alternate/reference expression ratio in autosomes as the null expectation.

Our analyses indicated an overall trend of reduced Y expression relative to X-linked alleles for both old and young categories (and similar results were obtained in a comparison with the full set of genes from the XY system; *SI Appendix*, Fig. S3), with the effect being markedly stronger for older Y-linked genes (median = 0.79; Wilcoxon test, $W = 1093796$, $P << 10^{-10}$; Fig. 2) than for the younger category (median= 0.90; $W = 495511.5$, $P = 0.0267$; Fig. 2 and *SI Appendix*, Fig. S3). The overall pattern suggests that Y-linked genes that spend more time in the nonrecombining regions are more likely to show functional deterioration. However, it is also possible that X-linked alleles have been up-regulated to some extent in males [partial dosage compensation (34, 35)], thus contributing to the observed lower Y expression relative to X (see below). Our results also suggest that some genes have elevated Y-linked expression relative to X-linked alleles (Fig. 2), although this is less common. The fact that younger sex-linked genes also show a significant reduction in their Y/X ratio indicates that reduced Y expression is probably one of the initial changes that occurs following the evolution of X-Y recombination suppression.

Disruption of normal expression levels and gene loss could negatively affect the fitness of males, potentially leading to selective pressure to up-regulate X-linked genes, a process known as dosage compensation (2, 34). To investigate this, we analyzed the expression of X-linked genes that were ascertained to be hemizygous in males (but present in two active copies in females) to determine whether such genes were hyperexpressed in males. Our analysis of 119 hemizygous genes revealed that relatively few hemizygous genes in males show evidence for a compensatory increase in gene expression compared with X-linked genes in females. The majority of X-linked genes with missing Y copies in males were expressed approximately twofold lower compared with females (Fig. 3*A*). In particular, a high proportion of these genes [94 (79%) of 119 genes] showed significantly lower expression in males than in females (*SI Appendix*, Table S6), whereas only 7 (6%) of 119 had significantly higher expression in males compared with one-half of total (X + X) expression in females. This suggests that dosage compensation is incomplete in *R. hastatulus*, and is evidently not mediated by a chromosome-wide mechanism that affects all X-linked genes similarly.

In contrast, we did not find a consistent reduction in male-specific expression for either old (Fig. 3*B*) or young (Fig. 3*C*) X/Y genes (*SI Appendix*, Table S6) compared with total X expression in females. This implies that the observed loss of expression of Y-linked alleles does not cause total levels of sex-linked gene expression in males to be reduced, potentially reflecting

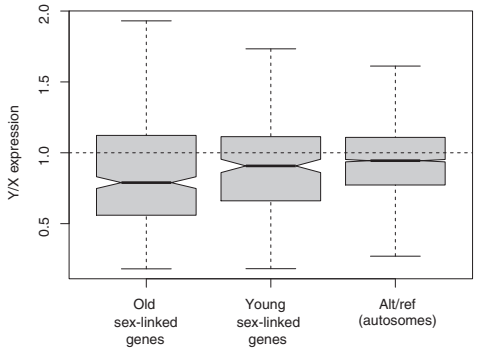

Fig. 2. Y/X gene expression of old and young sex-linked genes in *R. hastaulus*. The Y/X expression ratio distribution in males for 230 young sex-linked genes from the XY_1Y_2 system (not shared with the XY system) and 459 old sex-linked genes (shared with the XY system), compared with the expression ratio for alternate-to-reference (alt/ref) alleles at heterozygous sites in autosomes, is shown. Relative expression of Y alleles relative to X alleles was estimated per gene in males (i.e., within individual samples) by counting the numbers of mRNA reads covering sex-linked SNPs in sex-linked genes, and these relative estimates were averaged across all males. Expression estimates for reference and alternative alleles at heterozygous sites in autosomes were obtained similarly, using the numbers of mRNA reads covering SNPs in contigs where at least four such SNPs segregated as autosomal. The dotted line shows the expectation when X and Y alleles (or ref and alt alleles in autosomes) are equally expressed. Error bars show 1.5× the interquartile range, approximately corresponding to 2 SDs, and notches correspond approximately to 95% confidence intervals for the medians.

up-regulation of the male X allele to compensate for the loss in expression of the Y allele. However, it is unclear whether this compensatory increase in expression of the X allele in males is adaptive and was selected because of a degenerating Y allele. Instead, it may have arisen as a consequence of existing mechanisms of gene expression regulation that are activated in the presence of small perturbations in expression or gene dosage (e.g., refs. 36–38).

One potential complication of this analysis might be that changes in gene dosage on the sex chromosomes have led to sex-specific changes in autosomal expression, causing normalized estimates of male X-linked expression to be artificially deflated. To test whether there were global differences in autosomal expression between the sexes, we plotted the distribution of average expression in males divided by average expression in females for autosomal genes (39, 40) (*SI Appendix*, Fig. S7). This distribution is centered at 1 (n = 1,167, median = 1.01, SD = 0.349), suggesting a lack of widespread expression differences in males compared with females. A slight secondary peak around 1.9 is evident, suggesting that some genes may be differentially expressed, but the effect on the central tendency of the distribution is minimal. Although the slight right skew might mean that X up-regulation in males has been underestimated, we did not find evidence for large quantitative differences in autosomal expression, suggesting that our RNAseq-based estimates of X expression are reliable. Indeed, autosomal genes have the lowest level of differential gene expression between males and females (Fig. 3*D* and *SI Appendix*, Table S6), suggesting that most of the differential gene expression between the sexes is driven by sex chromosome evolution. Results consistent with this were obtained when examining expression differences in the XY system, as well as from independent population samples (*SI Appendix*, Table S6). Overall, we conclude that the majority of hemizygous genes are not dosage-compensated, whereas genes with retained Y copies have lower Y expression but no overall differential expression between the sexes.

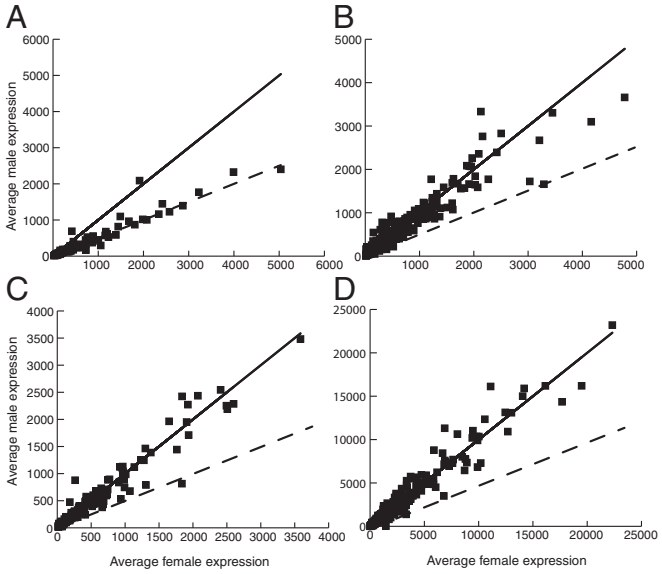

Fig. 3. Average normalized gene expression in male vs. female progeny (six of each sex) from the XY_1Y_2 system. Hemizygous genes (*A*), sex-linked genes with Y homologs shared with the XY race (old) (*B*), sex-linked genes with Y homologs not shared with the XY system (*C*, young), and autosomal genes (*D*) are shown. The solid line shows the expectation under equal male and female expression, and the dashed line shows the expectation for male expression being equal to one-half of female expression. Median differential expression normalization was conducted using DESeq (details are provided in *Methods*).

Molecular Evolutionary Tests for Deleterious Mutations and Codon Use Bias. We also tested whether the efficacy of purifying selection was reduced for Y-linked genes, and whether they have accumulated more deleterious mutations or changes in codon use compared with X-linked genes. This is expected because of the lower rate of recombination for Y-linked genes, which is predicted to reduce the efficacy of purifying selection (30, 31). However, given that recombination suppression was recent for many sex-linked genes, extensive deterioration of Y-linked genes may not be expected. Using our phased X and Y sequences, we used two approaches to test whether Y-linked sequences have accumulated deleterious changes. First, we used parsimony to estimate the total number of changes across sex-linked genes on the X and Y lineages, using orthologous sequences from *R. bucephalophorus*. The number of synonymous changes on the X vs. the Y for the old gene set is nearly equal, providing no evidence for elevated mutation rates on the Y chromosome (Fig. 4*A* and *SI Appendix*, Fig. S4). In contrast, nearly twice as many nonsynonymous changes have occurred on the Y lineage (1,646 vs. 835), implying reduced selection efficacy since the suppression of recombination. This difference is highly significant (Fisher's exact test, $P < 0.001$). For the young gene set, a weaker trend was apparent (339 vs. 215 nonsynonymous changes on the Y vs. the X; Fisher's exact test, $P < 0.001$; Fig. 4*A*).

We also generated maximum-likelihood estimates of ω, the nonsynonymous (dN) to synonymous (dS) substitution ratio for each lineage, including the X and Y sequences of both systems and the outgroup. Consistent with the parsimony approach, we found that old Y-linked genes in the XY_1Y_2 system had a higher number of nonsynonymous relative to synonymous substitutions per site compared with X-linked genes (average $\omega_{Y\_old} = 0.401$ and average $\omega_{X\_old} = 0.156$; Wilcoxon test, $P << 10^{-10}$), but the difference was much less and not significant for younger Y-linked genes (average $\omega_{Y\_young} = 0.209$ and $\omega_{X\_young} = 0.145$; $P = 0.114$). No significant difference in synonymous substitution rate was observed between X and Y chromosomes (Fig. 4*A* and *SI Appendix*, Fig. S4), suggesting that differences in ω are not due to differences in underlying mutation rates. Further, we found that old and young X sequences did not have significantly different ω values ($P < 0.399$), but the comparison of old vs. young Y genes revealed a significant difference ($P < 4 \times 10^{-8}$). As expected, analysis of substitution rates in the XY chromosome system gave comparable results to the old gene set in the XY_1Y_2 system (*SI Appendix*, Table S7). Together, these results indicate that elevated ω_{Y_old} is not due to changes on the X but is caused by a significantly higher substitution rate on the Y.

Finally, we also tested whether Y-linked genes have undergone more changes toward unpreferred codons than X-linked genes. Here, we used a parsimony approach to examine changes in codon use along the X vs. Y lineages, using the outgroup sequence to polarize changes on X and Y branches. To count the number of changes from preferred to unpreferred codons, and vice versa, we assumed shared codon preferences from *A. thaliana* (41). Old Y-linked genes had significantly more preferred-to-unpreferred changes in codon use relative to unpreferred-to-preferred changes compared with X-linked genes (Fig. 4*B* and *SI Appendix*, Fig. S5; Fisher's exact test, $P < 0.01$). However, no significant difference was observed in the ratio of codon changes for the young Y-linked genes (Fisher's exact test, $P > 0.05$). The larger number of codon substitutions in the old Y-linked genes may reflect a greater reduction in the efficacy of selection on codon use; additionally, differences in biased gene conversion due to recombination suppression may play a role. Collectively, these molecular evolutionary comparisons of X- and Y-linked sequences support the hypothesis that deleterious changes are accumulating in Y lineages as a result of a reduction in the efficacy of selection, with the magnitude of the effects depending on the time since recombination suppression.

Conclusions

Our segregation-based analysis using RNAseq has led to the identification of hundreds of sex-linked genes in a nonmodel

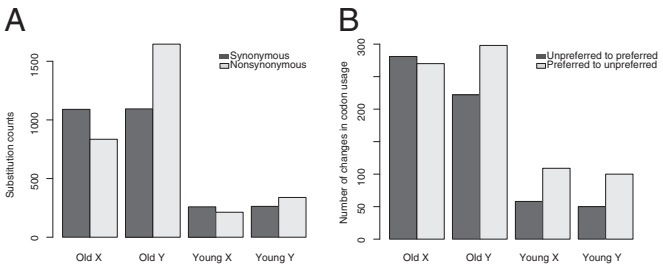

Fig. 4. Synonymous and nonsynonymous substitutions in X and Y genes. The number of parsimony-estimated lineage-specific substitutions (*A*) and changes in codon use (*B*) on the X and Y sequences from the XY_1Y_2 system are shown, using orthologous sequences from *R. bucephalophorus* to polarize changes along the X and Y lineages separately. Old genes represent those shared with the XY system, whereas young genes represent those that are not shared.

dioecious plant species with a neo-Y sex chromosome system. This has allowed us to compare the changes in expression and sequence evolution that have occurred following recombination suppression between X and Y chromosomes. The majority of X/Y genes in *R. hastatulus* have become nonrecombining recently and exhibit low X-Y sequence divergence; however, the older Y-linked genes that are shared between the XX/XY and XX/XY_1Y_2 systems show clear signs of degeneration, and many of the oldest sex-linked genes are likely in our hemizygous set. The older Y-linked genes have undergone gene loss, are accumulating nonsynonymous substitutions likely to impair gene function, contain more unpreferred changes in codon use, and show a loss of expression compared with X-linked genes. In contrast, we find that these features of Y degeneration are either significantly reduced or absent in the younger X/Y genes unique to the XX/XY_1Y_2 system. Our contrast between young and old sex-linked genes, made possible because of the unusual occurrence in *R. hastatulus* of intraspecific polymorphism in the sex chromosome system, provides a unique glimpse into the early stages and chronology of Y-chromosome degeneration in a flowering plant.

Methods

RNA Sequencing. To identify sex-linked genes in *R. hastatulus*, we sequenced transcriptomes from parents and F_1 progeny from two within-population crosses, one from a population with XY males (Many, LA; LA-MAN) and one from a population with XY_1Y_2 males (Branchville, SC; SC-BRA). We extracted RNA from leaf tissue using Spectrum Plant Total RNA kits (Sigma-Aldrich), and the isolation of mRNA and cDNA synthesis was conducted according to standard Illumina RNAseq procedures. Sequencing was conducted on the Illumina GAII platform for XX/XY parental samples with 80-bp end reads at the Center for the Analysis of Genome Evolution and Function (University of Toronto) and on the Illumina HiSeq platform by the Genome Quebec Innovation Center (GQIC) with 150-bp end reads for XX/XY_1Y_2 parental samples. F_1 samples were sequenced by multiplexing and barcoding six male and six female samples from each cross on a separate Illumina HiSeq lane with 150-bp end reads at the GQIC. Samples used for validation (see below; SNP segregation analysis and ascertaining sex linkage) were sequenced by barcoding and multiplexing on an Illumina HiSeq lane with 150-bp end reads at the GQIC. We also obtained 150-bp end RNAseq data for the transcriptome of one *R. bucephalophorus* individual from Spain, which was also sequenced at the GQIC with 150-bp end reads. This species has no sex chromosomes and was used as an outgroup.

Assembly of *R. hastatulus* Transcriptomes. We assembled a reference transcriptome de novo using Velvet [version 1.2.07 (42)] and Oases [version 0.2.08 (43)] and pooled paired end reads from six F_1 females of the XY_1Y_2 system. Using this as the reference transcriptome facilitated identification of sex-linked genes shared between the XY and XY_1Y_2 systems (as discussed in the next section). Before assembly, we trimmed the data to remove reads <50 bp, and VelvetOptimizer (version 2.2.4) was used to choose the best k-mer size for each individual transcript. To avoid missing low-coverage transcripts, the final total number of bases in each assembly was used to evaluate the best k-mer size, which was 43. Oases (version 0.2.08) was then run under default parameters. For each set of transcript isoforms, the longest was chosen as the final transcript. This reference assembly yielded 38,828 contigs (N50 = 2,089, total length = 44,585,937 bp). For the outgroup *R. bucephalophorus,* the assembly was run with the same pipeline, yielding a best k-mer length of 43 and 35,525 contigs (N50 = 1923, total length = 38,120,382 bp).

SNP Segregation Analysis and Ascertaining Sex Linkage. To assign sex linkage to assembled contigs in which nucleotide variants were identified, we mapped reads from both XX/XY and XX/ XY_1Y_2 samples to the reference transcriptome, assembled using reads from females of the XY_1Y_2 system. We conducted mapping using the Burrows–Wheeler Aligner [release 0.6.2-r126 (44)], followed by Stampy [release 1.0.20 (45)] for mapping more divergent reads. We used Picard tools (release 1.78, http://picard.sourceforge.net) to modify mapping output into the format required for the Genome Analysis Toolkit [GATK, version 2.1-11 (46)] variant calling software. We then conducted segregation analysis on both systems separately (*SI Appendix*) to obtain the set of sex-linked genes shared between the XY and XY_1Y_2 systems (referred to as the old sex-linked genes) and those that were unique to the XY_1Y_2 system (referred to as young sex-linked genes). The number of sex-linked genes identified as a function of the number of diagnostic polymorphisms is shown in *SI Appendix* (Table S1) for each system, along with the number shared between them. We required contigs to have four or more high-quality (Phred-scaled SNP quality score >60) SNPs, with genotype calls made for all parents and progeny from both sex chromosome systems and segregation patterns indicating sex linkage. Such sex-linked SNPs were identified based on either (*i*) the presence of a segregating Y-linked variant, where fathers and sons were heterozygous but mothers and daughters were homozygous, or (*ii*) the presence of a segregating X-linked variant, where fathers and daughters were heterozygous but mothers and sons were homozygous. To ensure that such X/Y contig assignments were reliable, we further filtered putative sex-linked contigs to include only those in which a segregating Y-linked variant was ascertained and showed the expected sex-specific genotypes in 12 population samples (*SI Appendix*). Such sites represent fixed differences between the X and Y. Similar approaches were used to identify hemizygous and autosomal genes (*SI Appendix*). All data parsing was done using Bash, R, or Perl. Scripts are available on request.

Comparisons of Sex-Linked Gene Expression. The number of mRNA reads covering sex-linked SNPs in sex-linked contigs was counted from the SNP output from GATK to obtain estimates of the relative expression of X- and Y-linked alleles in males. This enabled us to compare young and old sex-linked genes by determining their respective Y/X expression ratio distributions (Fig. 2). Because the relative expression of X and Y alleles was estimated per gene in males (i.e., within individual samples), it is unnecessary to normalize the counts across samples, and these relative estimates were averaged across all males. Expression estimates for reference and alternative alleles at heterozygous sites in autosomes were obtained similarly using the numbers of mRNA reads covering SNPs across all samples in contigs where at least four such SNPs segregated as autosomal. For gene-level (rather than allele-specific) expression comparisons of sex-linked and autosomal genes across the sexes, we estimated expression in coding sequences using HTSeq (47) with the "intersection-nonempty" option. We focused on coding sequences and excluded putative untranslated regions due to observed high variance in read counts in these regions. Following HTSeq, we used DESeq (48) to conduct median differential coverage normalization and test for differential expression using the beta binomial distribution. Genes with a maximum total read count across samples <20 were removed to eliminate loci with little power to test for differential expression. The possibility of widespread chromosome-wide differences in gene expression may complicate normalized expression tests in this system; however, we found that normalization using just autosomes gave nearly identical results, with no consistent bias by sex (*SI Appendix*, Fig. S6). Significant expression differences between the sexes were assessed using both a 5% cutoff and a 10% false discovery rate correction (*SI Appendix*, Fig. S6).

Consensus Contigs for Molecular Evolutionary Analysis. To analyze the molecular evolution of sex-linked genes, we generated X and Y consensus sequences based on parent and progeny genotypes using a phasing algorithm implemented in an R script (available upon request). For each nucleotide position within candidate sex-linked loci, we used sequencing coverage/quality information from parental samples to call sites that were identical on both X and Y copies. Sites were accepted as identical if both parental strains were called as homozygous and both had eightfold or greater sequencing coverage and genotype quality scores ≥60. Otherwise, sites were annotated as missing data. Candidate X/Y variants were initially identified as sites homozygous in the female parent and heterozygous in the male parent. Our

method used a likelihood ratio approach to evaluate the relative support for the heterozygous site representing a true X/Y variant (male: X^AY^a, female: X^AX^A) vs. a segregating X variant in the male (male: X^aY^A, female: X^AX^A). To test the performance of this method, we implemented a simulation that calculated likelihood ratio tests for simulated parent/progeny genotype arrays in which variants were either heterozygous X variants in the male parent or true X/Y variants (*SI Appendix*, Figs. S1 and S2).

ORF Identification, Sequence Alignment, and Phylogeny Reconstruction. We identified ORFs from consensus sequences for all X and Y consensus sequences and from orthologous *R. bucephalophorus* sequences (identified using a three-way reciprocal BLAST of contigs from each sex chromosome system plus the outgroup) using the "getorf" program from the EMBOSS suite (version 6.3.1) (49). For each locus, the X and Y ORFs from the XX/XY and XX/XY_1Y_2 systems, as well as the orthologs from the outgroup sequence, were aligned using MUSCLE (version 3.8.31) (50). We used ORF alignments to guide nucleotide alignments with in-frame gaps using a custom Perl script (available upon request). Maximum likelihood phylogenetic trees were produced from each nucleotide alignment using RaXML (version 7.0.4) (51).

Analysis of Evolutionary Rates. We used phylogenies as starting trees for the analysis of evolutionary rate at synonymous and nonsynonymous sites using PAML (version 4.6) (52). For each locus, we fit a "free-ratio" model (model = 1), allowing dN/dS to vary across branches. Branch-specific silent site divergence, dN/dS ratios, and tree topologies were then extracted and analyzed in R using the "phytools" package (53). For loci in which X and Y sequences were monophyletic across the sex chromosome systems, we estimated dN/dS as the average of the population-specific and ancestral branches, weighted by the corresponding dS values. For all other comparisons, only values at terminal branches were considered. For each alignment, we also used a modified version of Polymorphorama (54) to count the number of parsimony-estimated lineage-specific changes (synonymous, nonsynonymous, preferred→unpreferred, unpreferred→preferred) on the X and Y sequences, using the outgroup sequence to polarize changes. We analyzed the two sex chromosome systems separately for this analysis in a three-way alignment of X, Y, and outgroup.

ACKNOWLEDGMENTS. We thank Deborah Charlesworth for helpful advice, discussion, and comments; María Talavera Solís for seeds of *R. bucephalophorus*; and two anonymous reviewers for their comments on an earlier version of the manuscript. This research was funded by Natural Sciences and Engineering Research Council of Canada Discovery grants (to S.C.H.B. and S.I.W.).

1. Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* 355(1403):1563–1572.
2. Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* 6(2):149–162.
3. Bachtrog D (2013) Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* 14(2):113–124.
4. Liu Z, et al. (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* 427(6972):348–352.
5. Filatov DA (2005) Evolutionary history of *Silene latifolia* sex chromosomes revealed by genetic mapping of four genes. *Genetics* 170(2):975–979.
6. Yin T, et al. (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res* 18(3):422–430.
7. Spigler RB, Lewers KS, Main DS, Ashman TL (2008) Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity (Edinb)* 101(6):507–517.
8. Peichel CL, et al. (2004) The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Curr Biol* 14(16):1416–1424.
9. Kaminker JS, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: A genomics perspective. *Genome Biol* 3(12):H0084 Research 0084.1–0084.20.
10. Carvalho AB (2002) Origin and evolution of the *Drosophila* Y chromosome. *Curr Opin Genet Dev* 12(6):664–668.
11. Carvalho AB, et al. (2003) Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: How far can we go? *Genetica* 117(2-3):227–237.
12. Hughes JF, et al. (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463(7280):536–539.
13. Hughes JF, et al. (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483(7387):82–86.
14. Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286(5441):964–967.
15. Ross MT, et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434(7031):325–337.
16. Bergero R, Charlesworth D (2011) Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr Biol* 21(17):1470–1474.
17. Chibalina MV, Filatov DA (2011) Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol* 21(17):1475–1479.
18. Gschwend AR, et al. (2012) Rapid divergence and expansion of the X chromosome in papaya. *Proc Natl Acad Sci USA* 109(34):13716–13721.
19. Wang J, et al. (2012) Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc Natl Acad Sci USA* 109(34):13710–13715.
20. Bachtrog D (2006) Expression profile of a degenerating neo-Y chromosome in *Drosophila*. *Curr Biol* 16(17):1694–1699.
21. Kaiser VB, Zhou Q, Bachtrog D (2011) Nonrandom gene loss from the *Drosophila miranda* neo-Y chromosome. *Genome Biol Evol* 3:1329–1337.
22. Zhou Q, Bachtrog D (2012) Chromosome-wide gene silencing initiates Y degeneration in *Drosophila*. *Curr Biol* 22(6):522–525.
23. Zhou Q, Bachtrog D (2012) Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science* 337(6092):341–345.
24. Navajas-Pérez R, et al. (2005) The evolution of reproductive systems and sex-determining mechanisms within *Rumex* (Polygonaceae) inferred from nuclear and chloroplastidial sequence data. *Mol Biol Evol* 22(9):1929–1939.
25. Smith BW (1964) The evolving karyotype of *Rumex hastatulus*. *Evolution* 18(1):93–104.
26. Yamada K, et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302(5646):842–846.
27. Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. *Genetics* 175(4):1945–1954.
28. Nam K, Ellegren H (2008) The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. *Genetics* 180(2):1131–1136.
29. Orr HA, Kim Y (1998) An adaptive hypothesis for the evolution of the Y chromosome. *Genetics* 150(4):1693–1698.
30. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269–294.
31. Charlesworth B (1978) Model for evolution of Y chromosomes and dosage compensation. *Proc Natl Acad Sci USA* 75(11):5618–5622.
32. Skaletsky H, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942):825–837.
33. Degner JF, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25(24):3207–3212.
34. Mank JE (2013) Sex chromosome dosage compensation: Definitely not for everyone. *Trends Genet* 29(12):677–683.
35. Muyle A, et al. (2012) Rapid *de novo* evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS Biol* 10(4): e1001308.
36. Malone JH, et al. (2012) Mediation of *Drosophila* autosomal dosage effects and compensation by network interactions. *Genome Biol* 13(4):r28.
37. Birchler JA (1979) A study of enzyme activities in a dosage series of the long arm of chromosome one in maize. *Genetics* 92(4):1211–1229.
38. Devlin RH, Holm DG, Grigliatti TA (1982) Autosomal dosage compensation in *Drosophila melanogaster* strains trisomic for the left arm of chromosome 2. *Proc Natl Acad Sci USA* 79(4):1200–1204.
39. Sun L, et al. (2013) Dosage compensation and inverse effects in triple X metafemales of *Drosophila*. *Proc Natl Acad Sci USA* 110(18):7383–7388.
40. Sun L, et al. (2013) Differential effect of aneuploidy on the X chromosome and genes with sex-biased expression in Drosophila. *Proc Natl Acad Sci USA* 110(41):16514–16519.
41. Wright SI, Yau CBK, Looseley M, Meyers BC (2004) Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol* 21(9):1719–1726.
42. Zerbino DRD, Birney EE (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
43. Schulz MHM, Zerbino DRD, Vingron MM, Birney EE (2012) Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28(8): 1086–1092.
44. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
45. Lunter GG, Goodson MM (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21(6):936–939.
46. McKenna AA, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
47. Anders S, Pyl PT, Huber W (2014) HTSeq—A Python framework to work with high-throughput sequencing data. *bioRxiv*, 10.1101/002824.
48. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
49. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.
50. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
51. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
52. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
53. Revell LJ (2012) Phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223.
54. Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* 17(12):1755–1762.

Supporting Information (SI) Appendix for PNAS article:

Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*

Josh Hough*, Jesse D. Hollister, Wei Wang, Spencer C.H. Barrett, and Stephen I. Wright

Department of Ecology and Evolutionary Biology, University of Toronto,
25 Willcocks St., Toronto, Ontario, Canada, M5S 3B2

*Corresponding author: Josh Hough; josh.hough@utoronto.ca

Identification of sex-linked genes with Y-linked homologues

To identify sex-linked genes from the two sex chromosome systems, we filtered our genotype data (both SNPs and indels from both races) and required 1) all individuals to have a genotype call at the site, and 2) the SNP quality score of the site to be >=60. After these filters, we obtained a total of 730,957 polymorphic sites (SNPs or indels).

We identified SNPs showing XY segregation patterns separately in the XX/XY system and the XX/XY_1Y_2 system. In total, this led to the identification of 10,420 sex-linked SNPs from 1383 genes in the former system, and 16,967 SNPs from 2839 genes in the latter. The number of sex-linked genes identified, as a function of how many diagnostic polymorphisms are required, is shown in Table S1 for each system separately, and also shared between them (where the 'shared' criterion requires both system to have the same minimum number of segregating SNPs). In general, regardless of SNP cutoffs, roughly 70% of sex-linked genes are also identified as such in the XX/XY_1Y_2 system, while about 40% of genes found in XX/XY_1Y_2 are shared with the XX/XY system. This is consistent with the neo-Y system having acquired many new sex-linked genes since the autosomal fusion.

Table S1: Number of sex-linked genes with Y homologues as a function of the minimum number of SNPs required to identify sex-linked genes

Minimum number of SNPs with XY segregation	XY system	XY_1Y_2 system	Shared sex- linked genes	Percent of shared sex-linked genes
1	1383	2839	1005	73/35
2	1033	2043	747	72/37
3	838	1599	592	71/37
4	698	1298	510	73/39
5	616	1065	451	73/42

Population screen

We used population data (12 males, 12 females of each system, with one male and one female from each of six populations per sex chromosome system) to validate the ascertained sex-linked genes from crosses. Of particular interest was to distinguish candidate pseudoautosomal loci, which show linkage in the cross but not the population, from genes that are definitively in the sex-linked region. Note however, that by requiring fixed differences between the X and Y this will also exclude very recent sex-linked loci

that have not had time to accumulate fixed differences. The results of the polymorphism analyses are shown in Tables S2 and S3.

Table S2: Polymorphism screen in XX/XY_1Y_2 population

Minimum Number of SNPs used for identification	Number of genes with sex-linked patterns from XX/XY_1Y_2 population data	Overlap with crossing data with same minimum number of SNPs	Overlap with crossing data for ≥ 4 supporting SNPs
1	1210	954 (79%)	679
2	891	726 (81%)	609
3	723	592 (82%)	548
4	606	493 (81%)	493
5	530	432 (82%)	450

Table S3: Polymorphism screen in XX/XY system

Minimum Number of SNPs used for identification	Number of genes with sex-linked segregation in XX/XY population data	Overlap with crossing data with same minimum number of SNPs	Overlap with crossing data with ≥ 4 supporting SNPs
1	1294	946 (73%)	611
2	1000	746 (75%)	595
3	828	626 (76%)	572
4	723	550 (76%)	550
5	639	495 (77%)	518

Identification of autosomal genes

To screen for autosomal genes, we identified SNPs where the mother was homozygous, father heterozygous, and at least one son and at least one daughter was heterozygous (i.e. both sons and daughters inherit a focal allele from the father). Because this is a much less stringent filter for genotypes than sex-linked SNPs (e.g. all males being heterozygous and all females being homozygous), this category is more susceptible to genotyping errors. Indeed, the average coverage and genotype quality scores are lower for this set of SNPs than for the sex-linked SNPs (e.g. one focal Texas female has an average coverage of 32 and genotype quality of 54 for autosomal SNPs, but average coverage of 47 and quality of 84 for sex-linked SNPs). Because a major use of the autosomal SNPs was to provide a further filter for sex-linked genes and normalize gene expression comparisons, we therefore filtered the autosomal SNPs, requiring all individuals to have a minimum genotype quality score of 50, and removing SNPs that showed significant departures from Mendelian expectation in their genotype ratios. Following filtering, we were left with 890 genes with at least 4 high-confidence autosomal SNPs in the XX/XY system, and 1195 with at least 4 high confidence autosomal SNPs in the XX/XY_1Y_2 system. The numbers of identified autosomal genes as a function of the minimum number of SNPs used in the cutoff is shown in Table S4.

Table S4: Identification of autosomal genes and filtering of sex-linked genes showing autosomal SNPs

Number of SNPs used as cutoff	Number of autosomal loci, XX/XY system	Number of autosomal loci, XX/XY_1Y_2 system	Number of autosomal genes in XX/XY system that overlap with XX/XY sex-linked genes (4 SNP cutoff)	Number of autosomal genes in XX/XY_1Y_2 system that overlap with XY_1Y_2 sex-linked genes (polymorphism validated, 4 SNP cutoff)
1	3909	4005	26	22
2	2289	2566	17	15
3	1382	1750	10	8
4	890	1195	7	6

Identification of putative hemizygous genes

To search for genes showing hemizygous segregation, we looked for two types of segregation patterns, where A and B represent alternative SNPs or indels:

a) Maternal genotype AA, paternal genotype called BB. All daughters AB, all sons called AA
b) Maternal genotype AB, paternal genotype called AA (or BB), some daughters AB, no sons heterozygous, the set of sons showing BOTH AA and BB calls

Because these hemizygous segregation patterns rely only on X-linked polymorphisms and will have fewer SNPs than divergent X-Y homologues, we reduced the stringency of our criterion for defining hemizygous genes shared between the sex chromosome systems. In particular, we defined hemizygous genes in the XX/XY_1Y_2 system that are shared with the XX/XY system as those with at least 4 supporting SNPs in the XX/XY_1Y_2 system, and at least 1 supporting hemizygous SNP in the XX/XY system. Furthermore, to estimate the percent of sex-linked genes that are hemizygous, we identified the number of XY sex-linked genes with an equivalent number of segregating X-polymorphisms. The number of hemizygous genes identified as a function of the minimum numbers of SNPs used in the cutoffs is shown in Table S5.

Table S5: Identification of hemizygous genes as a function of the minimum number of SNPs used in cutoff

Minimum number of SNPs showing hemizygous segregation	Hemizygous genes in XX/XY system	Hemizygous genes in XX/XY_1Y_2 system	Shared hemizygous genes	Number of XY genes with X-linked maternal segregation pattern, XX/XY system	Estimated percent hemizygosity XX/XY system	Number of XY genes with X-linked maternal segregation pattern, XX/XY_1Y_2 system	Estimated percent hemizygosity XX/XY_1Y_2 system
1	571	496	209	540	49	1140	30
2	246	262	158	481	34	1018	20
3	159	192	125	426	27	915	17
4	119	144	100	373	24	819	15
5	79	112	82	334	19	757	13

Final filtered gene sets used in molecular evolution and expression analysis
To exclude putatively pseudoautosomal loci and other genes possibly misidentified as sex linked, we filtered our sex-linked genes to those that showed at least one sex-linked SNP from polymorphism data and at least 4 segregating sex-linked SNPs in crossing data. Furthermore, to exclude genes that were possibly erroneously identified as sex linked and/or represent chimeric assemblies, we also excluded any sex-linked genes showing any autosomal segregation patterns. This led to the following gene sets for expression and molecular evolution analyses:

XX/XY_1Y_2 shared: for the XX/XY_1Y_2 system, we found a total of 460 genes that show XY segregation in the family data (>=4SNPs), are shared with the XX/XYY system (where the XX/XY_1Y_2 system has >=4 supporting polymorphisms), have no autosomal SNPs, and have at least one fixed difference between X and Y in the population data. NONE of these overlap with the set of hemizygous genes identified in the XX/XY_1Y_2 system.

XX/XY_1Y_2 unique: for XX/XY_1Y_2 samples we found a total of 231 genes that show XY segregation in the family data, NOT shared with the other system, have no autosomal SNPs, and have at least one fixed difference between X and Y in the population data. None of these overlap with the set of hemizygous genes identified in the XX/XY_1Y_2 system.

XX/XY: For the XX/XY system, we found a total of 585 genes that show XY segregation in the family data, have no autosomal SNPs and are polymorphism validated

Autosomal in XX/XY_1Y_2 system: After removing autosomal genes with signs of sex-linkage, we ended up with a total of 1167 confidently autosomal genes with >=4 supporting SNPs in this system

Hemizygous in XX/XY_1Y_2 system: after removing any genes with 1 or more autosomal or XY SNP segregation pattern, and removing those with less than 4 supporting SNPs we are left with 122 autosomal genes in this system.

Hemizygous in XX/XY system: after removing any genes with 1 or more autosomal or XY SNP segregation pattern, and removing those with less than 4 supporting SNPs we are left with 106 genes.

Generation of consensus X/Y sequences
To analyze the molecular evolution of sex-linked genes, we generated X and Y consensus sequences. Sequences were generated based on the parent/progeny genotype information, using a novel phasing algorithm implemented in an R script (available upon request). For each nucleotide position within candidate sex-linked loci, we used sequencing coverage/quality information from parental strains to call sites that were identical on both X and Y copies. Sites were accepted as identical if both parental strains were called as homozygous, and both had ≥ 8x sequencing coverage and genotype quality scores ≥60. Otherwise sites were annotated as missing data.

Candidate X/Y variants were initially identified as sites homozygous in the female parent but heterozygous in the male parent. Our method then utilized a likelihood-ratio approach to evaluate the relative support for the heterozygous site representing a true X/Y variant (male: X^AY^a, female: X^AX^A) vs. a segregating X variant in the male (male: X^aY^A, female: X^AX^A). This method evaluated the probability that a given male or female progeny was heterozygous based on the binomial density function in R:

```
P(Het) = dbinom(x2,(x1+x2),0.5)
```

where x1 was the read support for the reference allele and x2 was the read support for the alternative allele. The likelihood of homozygosity was calculated using the dbinom function:

```
P(Hom) = dbinom(x2,(x1+x2),e) + dbinom(x2,(x1+x2),1-e)
```

Where e was equal to 1/3 the overall error rate at homozygous reference sites (calculated from errors in progeny sequence data at sites with high confidence homozygous parental genotypes). The likelihood of all males (or females) being heterozygous was then

$$L_{Het} = \prod_{i=1}^{n} P(Het)_i$$

And the likelihood of all males (or females) being homozygous was

$$L_{Hom} = \prod_{i=1}^{n} P(Hom)_i$$

Where $P(Het)_i$ and $P(Hom)_i$ represented the above probability for the i-th male (or female). The relative support for heterozygosity vs homozygosity was evaluated using separate likelihood ratio tests for males:

$$LRT_m = 2*[\log(L_{Het}) - \log(L_{Hom})]$$

and females:

$$LRT_f = 2*[\log(L_{Het}) - \log(L_{Hom})]$$

Finally, we evaluated support for the site being an X/Y variant by calculating the difference:

$$LRT_d = LRT_m - LRT_f$$

The LRT_d statistic takes on large values when all males have strong support for being heterozygous and all females have strong support for being homozygous, as expected for true X/Y variants (male: X^AY^a, female: X^AX^A).

To test the performance of this method, we implemented a simulation that calculated LTR_d for simulated parent/progeny genotype arrays in which variants were either heterozygous X variants in the male parent or true X/Y variants. We simulated gene expression levels by drawing from an exponential distribution fitted to the average coverage among individuals and across genes in our transcriptome dataset. We allowed a uniform distribution of missing data among male and female progeny, and assigned total coverage of both alleles for single individuals by sampling from a Poisson distribution with mean equal to the randomly drawn expression level. Based on the individual expression level, we then sampled alleles according to a binomial distribution where the probability of sampling the alternative allele was 0.5 for heterozygotes and e for homozygotes. We simulated 10,000 true X/Y variants (male: X^AY^a, female: X^AX^A) and 10,000 male X variants (male: X^aY^A, female: X^AX^A). Figure S1 shows the simulated distribution of LRT_d based on each of the above scenarios.

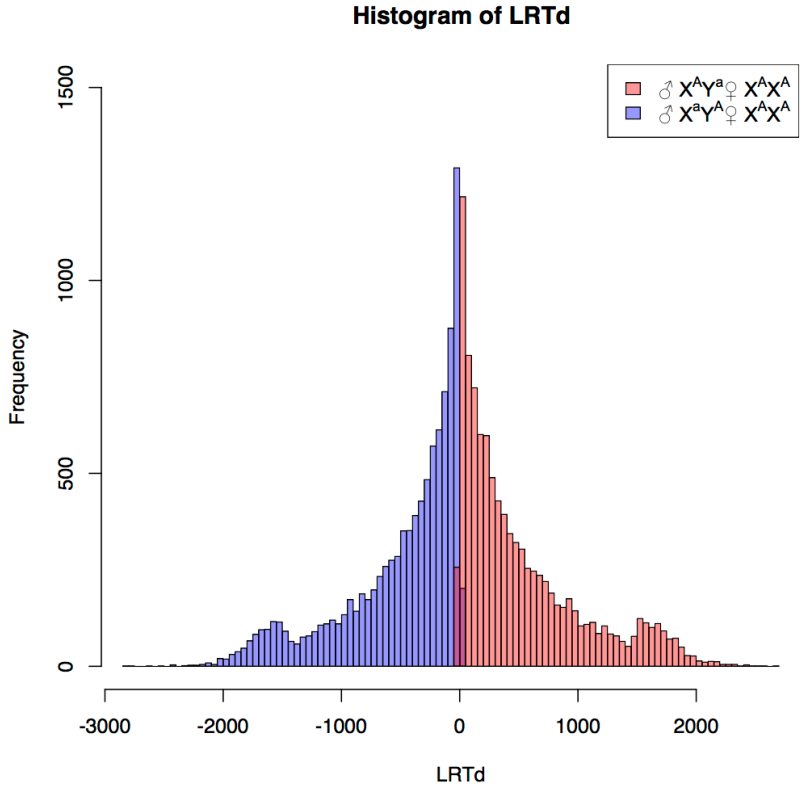

Figure S1: Histogram of the simulated distribution of the LRT_d statistic from 10,000 simulated parent/progeny SNP segregation patterns for either true X/Y variants (red) or a segregating X variant on the male X chromosome (blue). The distribution of LRT_d results from the strength of support for an X/Y variant in the presence of random sampling of gene expression level, read support for alleles, and proportion of missing data.

In addition, we quantified the Type I and Type II error rates for a given LRT_d cutoff, shown in Figure S2. We determined that an LRT_d cutoff of ≥4 would constrain the Type I error rate to <1% and the type II error rate to <5%. These are likely overestimates of the rates, however, as our simulation overestimated the occurrence of missing data by sampling that parameter from a uniform distribution.

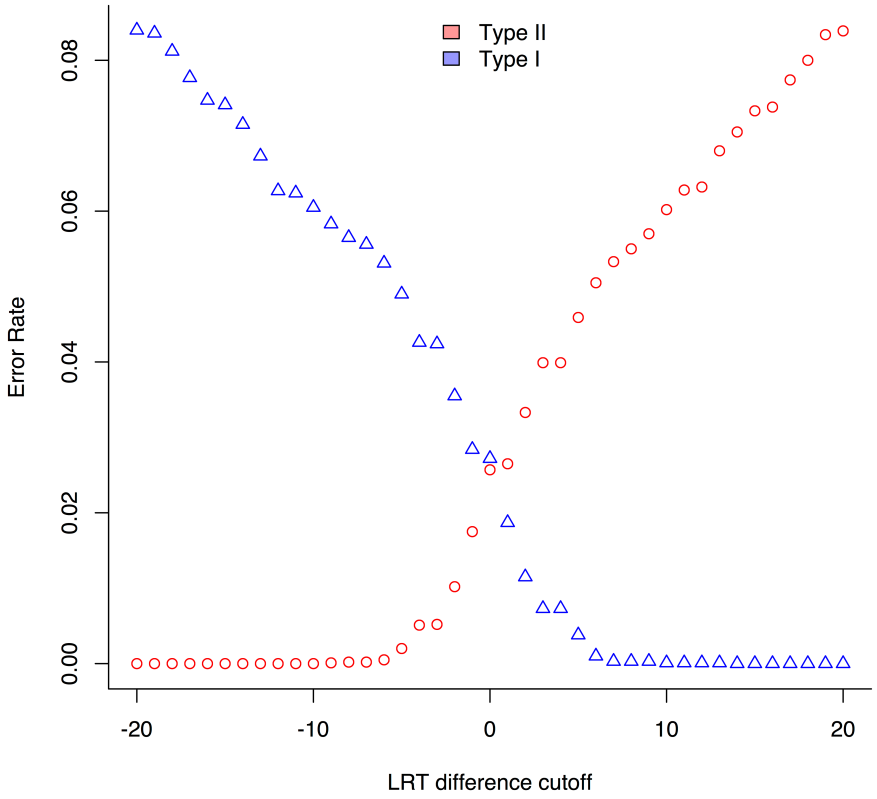

Figure S2. Plot of occurrence of Type I (blue) and Type II (red) error from 10,000 simulations of true X/Y or segregating X variants for values of the LRT_d statistic between -20 and 20.

Using this approach, consensus X/Y sequences were generated based on .vcf files separately for NC and TX populations. Because individuals from both populations were genotyped based on mapping to a North Carolina reference female, the North Carolina X/Y consensus sequences necessarily included a random sample of derived and ancestral states for segregating sites. To produce a similar outcome in the Texas sequences, we randomly assigned the non-reference base to the Texas X/Y consensus sequences 50% of the time. We identified putative fixed non-reference variants on the Texas X-chromosome based on a (male: X^aY^A, female: X^aX^a) pattern, and assigned them to the Texas X consensus if they had sufficient LRT_d support.

ORF identification, sequence alignment, and phylogeny reconstruction
Open Reading Frames (ORFs) were identified from consensus sequences for all X and Y consensus sequences, and from orthologous *R. bucephalophorus* sequences using the getorf program from the EMBOSS suite Version 6.3.1. For each locus, the X and Y ORFs from Texas and North Carolina, as well as the outgroup sequence, were aligned using MUSCLE Version 3.8.31. The ORF alignments were then used to guide nucleotide alignments with in-frame gaps using a custom PERL script (available upon request). Maximum likelihood phylogenetic trees were then produced from each nucleotide alignment using RaXML version 7.0.4.

Analysis of evolutionary rates
Phylogenies were used as starting trees for analysis of evolutionary rate at synonymous and non-synonymous sites using PAML Version 4.6. For each locus, we fit a "free-ratio" model (model=1), allowing dN/dS to vary across branches. Branch-specific silent site divergence, dN/dS ratios, and tree topologies were then extracted and analyzed in R using the "phytools" package. For loci in which X and Y sequences, respectively, were monophyletic across the two populations, we estimated dN/dS as the average of the population-specific and the ancestral branches, weighted by the corresponding dS values. For all other comparisons, only values at terminal branches were considered.

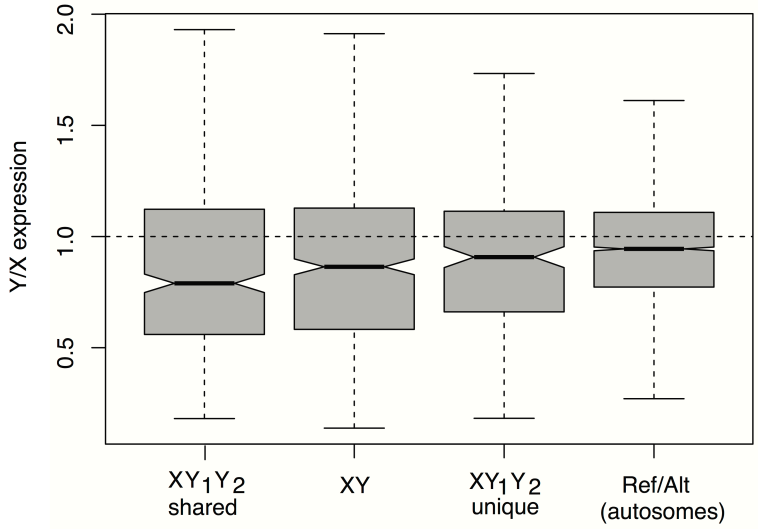

Figure S3. The Y/X expression ratio distribution in males for: 1) sex-linked genes from the XY_1Y_2 system shared with the XY system, 2) the full set from the XY system, and 3) unique to the XY_1Y_2 system, compared to the expression ratio for alternate to reference alleles at heterozygous sites in autosomes. Relative expression of X and Y alleles was estimated per gene by counting the numbers of mRNA reads covering sex-linked SNPs (or autosomal SNPs for autosomes). The dotted line shows the expectation when X and Y alleles (or ref and alt alleles in autosomes) are equally expressed. Error bars show 1.5 times interquartile range, approximately corresponding to two standard deviations, and notches correspond approximately to 95% confidence intervals for the medians.

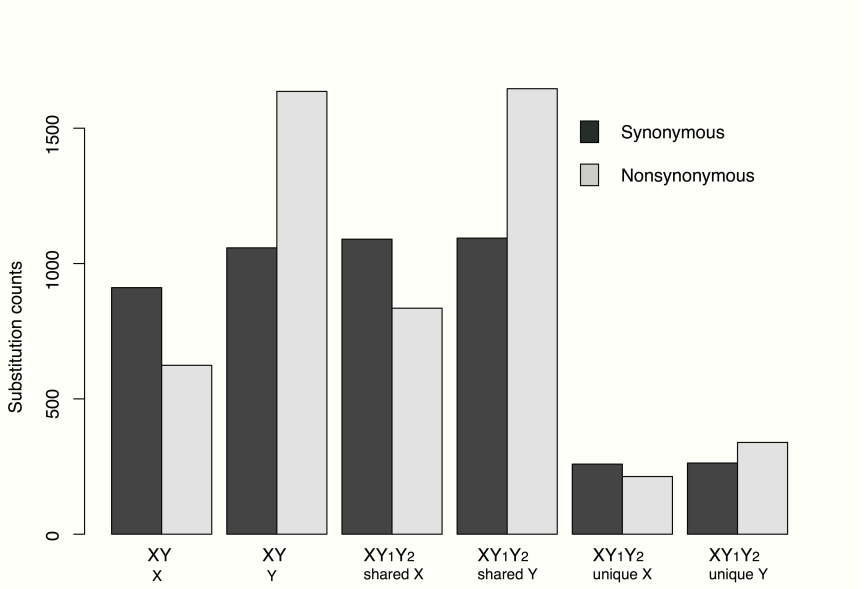

Figure S4. The number of parsimony-estimated lineage-specific substitutions (synonymous or nonsynonymous) on the X and Y sequences from the XY and XY_1Y_2 systems, using orthologous sequences from the outgroup, *R. bucephalophorus,* to polarize the changes along the X and Y lineages separately. 'Shared' genes represent those shared with the XY system, while 'unique' genes represent those that are not shared.

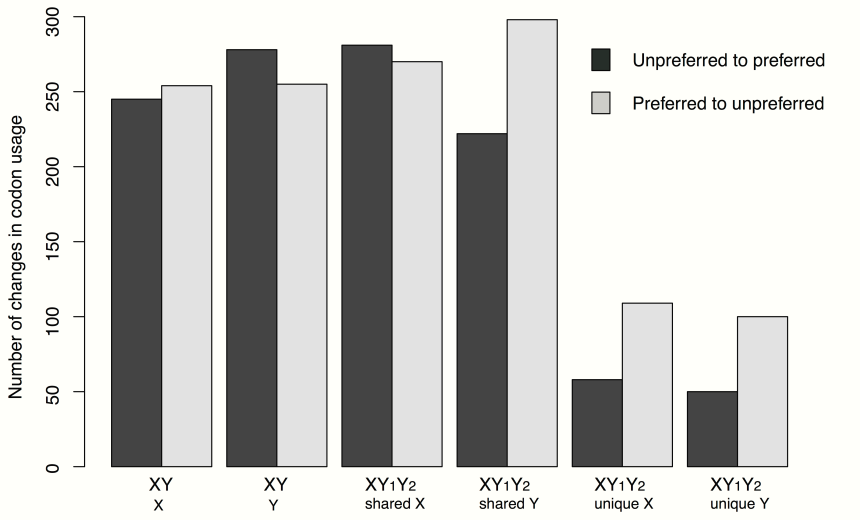

Figure S5. Changes in codon usage for X and Y genes. Total numbers of parsimony-estimated lineage-specific changes from preferred-> unpreferred and unpreferred->preferred for the XY and XY_1Y_2 systems. 'Shared' genes represent those shared with the XY system, while 'unique' genes represent those that are not shared.

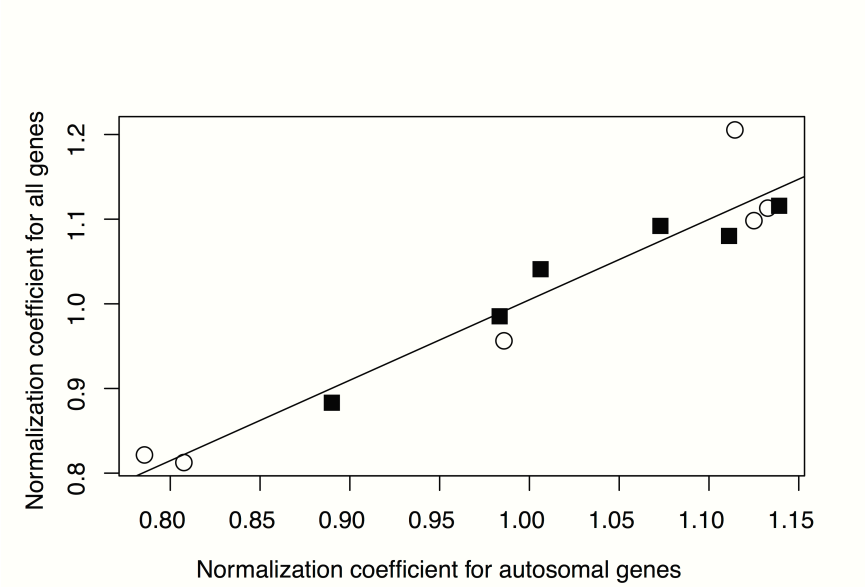

Figure S6. DeSeq normalization coefficients ('scaling factors') from all genes compared with normalization using just autosomal genes for the XY_1Y_2 system progeny data. Males, squares, and females, circles. No clear bias is observed using scaling factors from either the total gene set or autosomal genes alone.

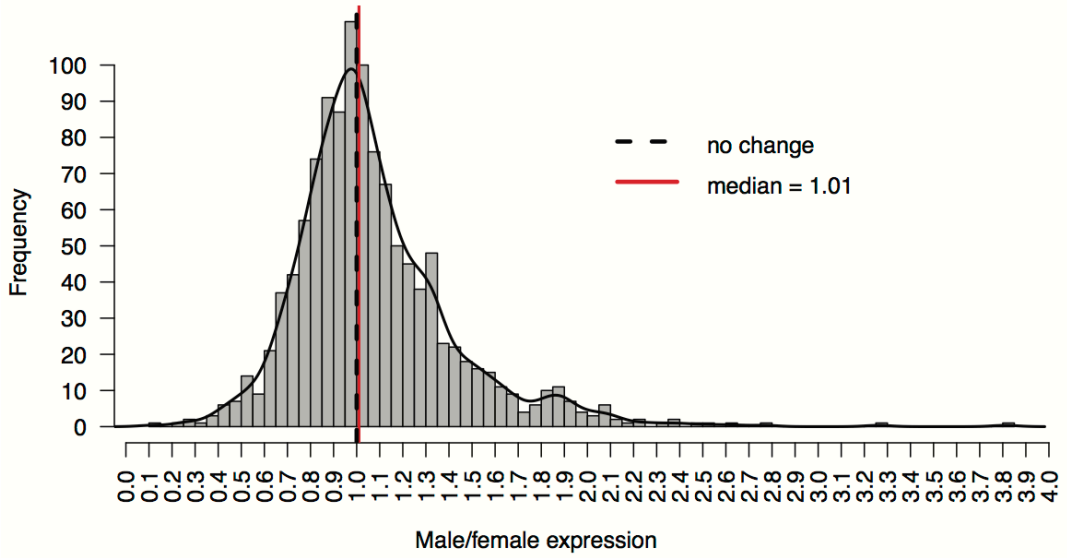

Figure S7. Distribution of average autosomal gene expression in males divided by average expression in females for the XY_1Y_2 system progeny data (6 males; 6 females; 1167 autosomal genes). The dotted line shows a male/female ratio of 1, indicating no sex-specific change. We normalized each sample by the total number of mapped reads to make the 12 biological replicates comparable.

Table S6: Results of statistical tests of differences between male and female expression for different gene sets. Statistical tests were performed throughout using the beta binomial test (see methods), with median differential expression normalization.

Gene set[1]	Total tested	Number significant, 5%	Number significant, 10% FDR	Percent significant 5%	Percent significant, 10% FDR	Number of genes significant at 5% that show female overexpression	Number of genes, 10% FDR that show female overexpression
Autosomal, XY_1Y_2 system, progeny	1167	32	2	2.7	0.2	9	1
Hemizygous, XY_1Y_2 system, progeny	119	94	47	79	39.5	94	47
'Old', XY_1Y_2 system, progeny	458	72	21	15.7	4.6	32	5
'Young', XY_1Y_2 system, progeny	167	15	6	9.0	3.6	6	2
Autosomal, XY_1Y_2 system, population	1166	179	48	15.4	4.1	62	14
Hemizygous, XY_1Y_2 system, population	119	92	52	77.4	43.7	91	52
'Old', XY_1Y_2 system, population	458	97	22	21.2	4.8	37	5
'Young', XY_1Y_2 system, population	167	31	8	18.6	4.8	15	3
'Old', XY system, population	584	64	5	11.0	0.9	25	0
Hemizygous, XY system, population	105	68	5	64.8	4.8	68	5
Autosomal, XY system, population	889	23	1	2.6	0.1	14	1
Hemizygous, XY system, progeny	104	57	24	54.8	23.1	55	24
'Old', XY system, progeny	578	123	39	21.2	6.7	48	15
Autosomal, XY system, progeny	888	156	46	17.6	5.2	59	17

1. Gene sets include male and female progeny from crosses ('progeny'), and data from 12 population samples ('population') in all cases six males were compared with six females using a global normalization procedure for all 12 samples. 'Old' genes in the XY_1Y_2 system represent those shared with the XY system, while 'young' genes are not shared. 'Old' genes for the XY system represent the entire complement of genes identified in this system. In all cases only genes with a maximum expression of at least 20 reads across samples were retained for analysis.

Table S7: Chromosome-specific PAML estimates of the per-site synonymous substitutions rate (Ks) and the ratio of nonsynonymous to synonymous substitutions (ω) in sex-linked genes.

Sex chromosome system	Gene set	Average Ks (Standard Error)	Average ω (Standard Error)
XY_1Y_2	Old X	0.00870 (0.00411)	0.156 (0.0379)
	Old Y	0.0120 (0.00160)	0.401 (0.0533)
	New X	0.00276 (0.00116)	0.145 (0.0553)
	New Y	0.00297 (0.00109)	0.209 (0.0730)
XY	Old X	0.00661 (0.00116)	0.169 (0.0367)
	Old Y	0.0122 (0.00185)	0.381 (0.0494)