Materials and Methods

1. Sequencing and Assembly

A highly inbred strain of *Drosophila miranda* (strain MSH22) was used for genome sequencing and RNA-seq analysis. Genomic DNA was extracted separately for pooled males and females (DNeasy Blood & Tissue Kit, Qiagen), sheared and size-selected. We prepared mate-pair libraries with an insert size of 210bp and sequenced 75bp reads from both ends following the standard manufacturer's instructions. We sequenced 3 lanes from both the male and female library, resulting in a total of 4.4 Gb and 5.0 Gb of high-quality reads, respectively (a total of over 60 million high-quality read pairs, i.e. ~30-fold coverage for each sex, implying 45-fold coverage for the X and neo-X, 15-fold coverage for the Y and neo-Y, and 60-fold coverage for autosomes).

High levels of sequence similarity between the neo-sex chromosomes and heterochromatic DNA in males deteriorate joint assemblies (table S1); thus we initially assembled male and female short reads separately into scaffolds using SOAPdenovo (http://soap.genomics.org.cn/soapdenovo.html, -K 31) (*26*). A high proportion of reads (>78%) participated in the assemblies, resulting in a genome size of about 125Mb and N50 lengths of 5.3kb and 13.8kb for male and female, respectively. We ordered as many scaffolds of *D. pseudoobscura* as possible into chromosomal sequences based on physical map information (*27*), in order to assign locations of *D. miranda* scaffolds. We aligned male and female *D. miranda* scaffolds against the chromosomal sequences of *D. pseudoobscura* using BLAST (*28*) (-e 1e-5), and grouped male and female scaffolds mapping to chromosomes other than chr3 (the neo-sex chromosome of *D. miranda*) together as the 'non-neo-sex' scaffolds, and the female scaffolds mapping to chr3 as 'neo-X' scaffolds. Short reads were then mapped to these two groups of scaffolds by SOAP (*29*), adjusting for mate-pair direction and insert-size information. We collected a total of 5.4 Gb non-neo-sex short reads from both sexes, and these pooled reads were used to re-assemble the non-neo-sex chromosomes of *D. miranda*, resulting in an improved assembly of autosomal and X chromosome scaffolds (with a N50 statistics of 23.7kb; i.e. 50% of the genome assembly is contained in scaffolds equal or larger to this value). (table S1).

To assemble the neo-Y chromosome, we need to distinguish reads derived from the neo-X and neo-Y from the male reads pool. We first separately used the male and female reads, and identified all the SNP sites along the neo-X scaffolds by SOAPsnp, with a cutoff of at least 3 uniquely-mapped supporting reads for both SNP sites (*30*). All the male reads containing male-specific SNPs were collected into the neo-Y reads pool, and those that mapped to the neo-X sequence at such sites were added to the neo-X reads pool. We also counted mapped reads number for each site of the neo-X scaffolds, and then calculated mapping coverage as the average reads count within 100bp sliding windows. We investigated the ratios of mapping coverage calculated from male reads versus female reads, in order to compartmentalize the divergent and non-divergent regions between neo-X and neo-Y chromosomes. The distribution of the ratio reveals two pronounced peaks centered at 1 and 0.5 (fig. S1), which correspond to the non-divergent and divergent regions between the neo-sex chromosomes. We defined all the regions with coverage lower than 0.5 as divergent regions. Reads mapping to these regions but not harboring any male-specific SNPs were removed from the male reads pool as neo-X reads. We combined all the remaining male reads (including all non-divergent regions) and neo-Y reads collected before, and performed a second round of *de novo* assembly for neo-Y scaffolds, as well as scaffolds from the neo-X and other chromosomes using reads of both sexes. Gaps within the scaffolds were finally filled by extending the mapped reads into the gap region by the Gapcloser module of SOAPdenovo.

We assessed the quality of our assemblies by comparing it to over 4-Mb of Sanger- and 454-derived BAC sequences (*16*, *31*). The integrity of the assembly was measured by BLAT alignment (using default parameters) as the ratio of summed aligned block lengths vs. target sequence length, while the accuracy was measured as ratio of identical cases vs. aligned lengths.

91.1% of neo-X and 65.1% of neo-Y BAC sequences collinearly aligned with the corresponding short-read assemblies, with alignment identities of 99.9% and 99.2%. Autosomal and X-linked BAC clones also show high alignment scores (90.5% aligned with 99.9% identity). Coding regions show similarly high alignment rate (99.1% vs. 94.9%) and identity (99.9% vs. 99.4%) for the neo-X and neo-Y. This suggests that lower alignment scores on the neo-Y mainly result from difficulties in assembling repetitive intergenic regions or structural variations (see fig. S2, table S2).

To further validate the assembly, and independently verify partly or entirely deleted genes from the neo-Y, we additionally sequenced males from the same *D. miranda* strain at medium genomic coverage with 454 technology. We produced 947,399 reads (maximum read length: 1,433-bp, mode read length 746-bp, median read length 667-bp), adding up to 587Mb (~5 fold genomic coverage) of data. Comparison of these raw reads to the Illumina assembly reveals a similarly high coverage (~90%) and identity (99.5%) on autosomes, and slightly lower similarities on chrXL, chrXR and the neo-sex chromosomes, due to lower sequencing coverage (table S3). We aligned male 454 reads back to the Illumina assemblies, to confirm partially or completely deleted neo-Y genes (fig. S3). Genes identified as having deletions from the neo-Y consistently show a length difference with the 454 raw reads (fig. S3A) or a lower male vs. female mapping coverage (fig. S3B), comparing to complete neo-Y genes.

2. Mapping and Annotation

To order the *D. miranda* scaffolds into chromosomes, we aligned the scaffolds against *D. pseudoobscura* chromosomes with BLAST (*2*) (-e 1e-5). Alignment hits of the same query scaffold with intervals shorter than 10kb were grouped together based on collinearity and screened with a 30% alignment length percentage cutoff. The best hit as well as others whose length percentages differ with that of the best one less than 10% were maintained as 'mapped' *D. miranda* scaffolds. Scaffolds aligned to an overlapping region of *D. pseudoobscura* with another longer scaffold were removed as redundancies. We then ordered and concatenated *D. miranda* scaffolds into chromosomes, filling gaps between them with the same size as the *D. pseudoobscura* region between the two alignment hits. Neo-X and neo-Y scaffolds were ordered separately into neo-X and neo-Y chromosomal sequences.

We aligned (TBLASTN, -e 1e-5) the protein sequences of *D. pseudoobscura* (www.flybase.org, v2.6) with the *D. miranda* chromosomal sequences to annotate *D. miranda* genes. Collinear alignment hits of the same protein query distant from each other with a length shorter than 10kb were chained together and the whole spanned genomic regions were taken as the candidate *D. miranda* gene regions. We finally used GeneWise (version 2.41) (*32*) and predicted open reading frames for *D. miranda* genes. We applied the annotation procedure to the neo-X and neo-Y chromosome separately. Putatively non-functional genes along the neo-Y chromosome were characterized as those with either premature terminal codons (PTC) or frame-shift mutations.

We validated PTCs in non-functional neo-Y genes by examining 454 reads aligned to these stop codon sites. Out of 118 neo-Y linked genes with PTC, 454 reads mapped (at least partly) to 81 genes (68.6%) in our neo-Y assembly, and 67 of these aligned copies (82.7%) have at least one 454 read supporting the PTC. In the remaining 41 cases, the read did not span the PTC or did not overlap other SNPs, preventing us to distinguish between the neo-X and neo-Y copy. Thus, despite the relatively low coverage of the 454 data, we could verify a large fraction of the PTC on the neo-Y.

We also compared our non-functional gene set to a list of 118 previously studied neo-sex gene pairs (70 functional, 48 non-functional) that were manually annotated using *D. pseudoobscura* gene models as a guide (*16*). We compared these data with our Illumina assembly and the 454 sequencing reads, and identified 5 cases where our Illumina assembly inferred a functional gene that was identified as a pseudogene in the previous study (*16*) and confirmed as such by 454 reads; these cases appear to be caused by assembly artifacts, where neo-X reads are incorporated into a neo-Y scaffold. In contrast, we never incorrectly annotated a functional gene as a pseudogene on the neo-Y when comparing the two assemblies. Thus, our false positive rate is 0 (calling a functional gene a pseudogene) and our false-negative rate (missing a pseudogene) is ~10%. On the neo-X, all but 2 genes were correctly annotated; one of the missing genes did not pass our sequence identity cutoff, and the second was in a region of poor assembly quality.

3. Evolutionary Analysis

We calculated SNP densities and read coverage every 50kb based on the neo-X chromosome as the reference, with a sliding window size of 5kb. The mapping and SNP calling procedures are the same as described above and results were plotted with *Circos* (*33*). We identified pairs of regions that failed to align between the neo-sex chromosomal sequences, with at least 100bp aligned at their flanking regions. These are divergent regions homologous to each other, and we then ran RepeatMasker (http://www.repeatmasker.org/) on regions with either of the pair longer than 500bp to compare their repeat contents (fig. S2A). We identified violations of pair-end mapping or insert distance (differ by at least 150bp) as signs of structural variations, using a similar algorithm as described in (*34*). For example, at a tandem duplication in the genome, read pairs spanning the duplicate boundary would be aligned as ‘reverse-forward’ rather than the normal ‘forward-reverse’ direction (*34*). A more stringent criterion of at least 5 informative read pairs were required for identifying a SV event.

For the gene ontology study, we classified *D. miranda* genes into GO categories, using information from their *D. melanogaster* orthologs and subjected them to enrichment analysis relative to the gene content of the entire genome using *Ontologizer* (*35*). We aligned coding sequences of neo-X, neo-Y and the orthologous *D. pseudoobscura* genes guided by the protein sequence using *TranslatorX* (*36*). We then used a likelihood method to calculate the pairwise ω (the ratio of amino-acid (*K*a) vs. synonymous site (*K*s) divergence) between *D. miranda* and *D. pseudoobscura* genes (*37*). Any genes with either *K*a or *K*s higher than 0.5 or ω value higher than 5 were excluded from further analysis, due to possible alignment artifacts or too few synonymous changes. To detect elevated evolution along specific branches, we used *codeml* of the *PAML* package and first assumed ω to be the same for the neo-X, neo-Y and outgroup branch (one-ratio model). The two-ratio model, allowing two different ω values (one for the branch of interest and another one for the background branches) was then compared to the one-ratio null model (*38*). Twice the log likelihood difference was compared with a $\chi 2$ distribution to test whether the data fits the two-ratio model significantly better (*39*). The free-ratio model, which allows different ω ratios for all the branches investigated, was used to calculate branch-specific evolutionary rates. Optimal codons for Drosophila were taken from reference (*40*).

4. **Expression analysis**

We prepared RNA-seq libraries from whole adult virgins of males and females, as well as dissected testes, accessory glands, male body carcasses, ovaries and female body carcasses from both *D. miranda* and *D. pseudoobscura*. Virgin flies were collected and aged for an extra week allowing for maturation before the dissection and RNA extraction (RNAeasy Tissue Kit, Qiagen). Poly(A)+ transcripts were isolated using Dynal magnetic beads (Invitrogen) and fragmented by ‘RNA fragmentation reagent’ (Ambion). Random primers were used to reverse transcribe the mRNAs and the mate-pair libraries with an insert size of about 250bp were prepared following the manufacturer’s standard protocol. Each sample has been sequenced on one lane and 75bp from both ends (2~4Gb data per sample).

We aligned the RNA-seq reads against the neo-X chromosome sequence with *TopHat* (*41*) and assigned individual reads as neo-X or neo-Y linked based on the pre-identified genomic SNP information. Neo-X or neo-Y reads spanning diagnostic SNPs within a gene were summed together to measure the allelic expression level. We used logarithmic-scaled ratios of neo-X read number vs. neo-Y read number to measure the degree of neo-X biased expression. We conservatively set the allelic read counts as 10 when they are lower than 10, for performing the binomial test of significance of biased gene expression. The expression levels of genes were calculated as FPKM (fragments per kilobase of exon per million fragments mapped) as defined by *Cufflinks* (*42*). We ranked expression level of genes along each chromosome by their FPKM values, and investigated chromosomal distributions of top 100, 200, 500, 1000, 2000 genes. The expected numbers of top expressed genes on each chromosome were calculated using total gene number of that chromosome assuming an even distribution. Patterns of demasculinization do not change for different numbers of genes investigated and only the pattern of top 500 genes is shown. For the regression analysis, we used expression levels from *D. pseudoobscura* as a proxy for ancestral expression level of *D. miranda* genes prior to the formation of the neo- X chromosome, so that we could exclude possible demasulinization effects on gene expression. We used logarithmic-scaled FPKM values and ω ratios, and excluded genes without any detectable expression or nonsynonymous changes from the analysis.

table S1 Assembly statistics using male, female or pooled reads

	Scaffold N50 (bp)	Average Scaffold Length (bp)	Scaffold Numbers	Longest Scaffold (bp)	Assembled Size (Mbp) incl. Ns	Reads Participated (%)
♂	5312	2289	52580	128191	120	78.2
♀	13773	5637	22259	102796	125	82.7
♂+♀	5007	2376	47035	41135	112	69.8
neo-X	23056	7112	2659	89551	20	NA
neo-X*	27677	18533	832	89551	15.4	NA
neo-Y	715	476	36282	10995	22	NA
neo-Y*	2298	1880	1962	10995	3.7	NA

* Assembly statistics using only scaffolds that contain coding regions

table S2 Structural variation numbers on each chromosome identified by male or female reads

Male/female	chrXL	chrXR	chr2	neo-X/Y	chr4	chr5
Deletion	31/56	43/73	96/89	134/51	72/78	5/0
Dispersed Duplication	2/3	3/5	8/7	7/7	10/11	0/0
Tandem Duplication	14/16	24/29	39/35	24/15	37/39	0/0

table S3 Similarity between male 454 reads and the Illumina assembly

	chrXL	chrXR	chr2	chr4	neo-X	neo-Y
Aligned bp	15170000	21295168	25039899	28347019	17696342	18921553
Total length	20731836	29473195	27885022	31286013	20038917	22216834
Aligned	73.17%	72.25%	89.80%	90.61%	88.31%	85.17%
Identity (mean/median)	98.7%/ 100%	99.1%/ 100%	99.5%/ 100%	99.5%/ 100%	98.5%/ 98.9%	97.8%/ 98.7%

table S4 Genomic divergence between the neo-sex chromosomes

Average coverage (cvg.) and SNP density (dst.) calculated from male and female *D. miranda* reads for each chromosome are shown. Increased SNP density of the neo-sex chromosomes in males reflects divergence between neo-X/neo-Y chromosomes.

	chrXL	chrXR	neo-sex (chr3)	chr2	chr4	chr5
♂ cvg.	8.205	8.017	12.190	16.237	15.250	10.806
♀ cvg.	15.803	15.466	16.837	15.859	15.226	11.573
♂ SNP dst. (sites/kb)	0.071	0.045	3.696	0.161	0.075	0.125
♀ SNP dst. (sites/kb)	0.097	0.061	0.080	0.142	0.071	0.098

table S5 Enriched GO terms for genes with intact neo-Y ORFs

GO ID	Namespace	Name	P-value
GO:0007268	biological_process	synaptic transmission	0.003517419
GO:0016202	biological_process	regulation of striated muscle tissue development	0.009399237
GO:0048634	biological_process	regulation of muscle organ development	0.003859401
GO:0042052	biological_process	rhabdomere development	0.005932791
GO:0008300	biological_process	isoprenoid catabolic process	0.00979735
GO:0051231	biological_process	spindle elongation	0.004891947
GO:0006818	biological_process	hydrogen transport	0.003586818
GO:0009636	biological_process	response to toxin	1.58E-04
GO:0044087	biological_process	regulation of cellular component biogenesis	0.002196067
GO:0035149	biological_process	lumen formation, open tracheal system	0.005982934
GO:0007113	biological_process	endomitotic cell cycle	0.009360521
GO:0007173	biological_process	epidermal growth factor receptor signaling pathway	0.006608712
GO:0065007	biological_process	biological regulation	0.007005348
GO:0044237	biological_process	cellular metabolic process	0.005286921
GO:0006716	biological_process	juvenile hormone metabolic process	0.007905138
GO:0031989	biological_process	bombesin receptor signaling pathway	0.006974717
GO:0032501	biological_process	multicellular organismal process	2.23E-05
GO:0014070	biological_process	response to organic cyclic compound	0.003792779
GO:0007307	biological_process	eggshell chorion gene amplification	6.15E-04
GO:0015992	biological_process	proton transport	0.005379925
GO:0003014	biological_process	renal system process	0.002577951
GO:0019731	biological_process	antibacterial humoral response	0.00978717
GO:0000022	biological_process	mitotic spindle elongation	0.002597837
GO:0048519	biological_process	negative regulation of biological process	0.006761734
GO:0032502	biological_process	developmental process	0.00417023
GO:0000267	cellular_component	cell fraction	5.31E-04
GO:0005940	cellular_component	septin ring	0.002102785
GO:0044422	cellular_component	organelle part	3.96E-04
GO:0005839	cellular_component	proteasome core complex	0.004906022
GO:0019773	cellular_component	proteasome core complex, alpha-subunit complex	0.002425214
GO:0005623	cellular_component	cell	6.22E-04
GO:0031105	cellular_component	septin complex	2.33E-04
GO:0032156	cellular_component	septin cytoskeleton	3.90E-04
GO:0016469	cellular_component	proton-transporting two-sector ATPase complex	0.006140409
GO:0016604	cellular_component	nuclear body	4.42E-04
GO:0043226	cellular_component	organelle	9.28E-04
GO:0030532	cellular_component	small nuclear ribonucleoprotein complex	0.009501371
GO:0044464	cellular_component	cell part	6.22E-04

GO:0032991	cellular_component	macromolecular complex	2.70E-04
GO:0005372	molecular_function	water transmembrane transporter activity	1.51E-05
GO:0005217	molecular_function	intracellular ligand-gated ion channel activity	0.003863038
GO:0004972	molecular_function	N-methyl-D-aspartate selective glutamate receptor activity	0.001832845
GO:0020037	molecular_function	heme binding	0.002061428
GO:0004693	molecular_function	cyclin-dependent protein kinase activity	0.002582101
GO:0015250	molecular_function	water channel activity	7.63E-04
GO:0022891	molecular_function	substrate-specific transmembrane transporter activity	0.002521138
GO:0004946	molecular_function	bombesin receptor activity	0.006644518
GO:0070003	molecular_function	threonine-type peptidase activity	0.001580524
GO:0004175	molecular_function	endopeptidase activity	0.009983873
GO:0004298	molecular_function	threonine-type endopeptidase activity	0.004006073
GO:0022892	molecular_function	substrate-specific transporter activity	0.007549218
GO:0051864	molecular_function	histone demethylase activity (H3-K36 specific)	0.003658537
GO:0008324	molecular_function	cation transmembrane transporter activity	0.001146674
GO:0042625	molecular_function	ATPase activity, coupled to transmembrane movement of ions	8.37E-04
GO:0009055	molecular_function	electron carrier activity	5.72E-04
GO:0005549	molecular_function	odorant binding	0.001231431

table S6 Enriched GO terms for genes with disrupted neo-Y ORFs

GO ID	Namespace	Name	P-value
GO:0046834	biological_process	lipid phosphorylation	0.009405102
GO:0019637	biological_process	organophosphate metabolic process	0.00793824
GO:0046165	biological_process	alcohol biosynthetic process	3.86E-05
GO:0007165	biological_process	signal transduction	0.005367977
GO:0034660	biological_process	ncRNA metabolic process	0.002410273
GO:0016042	biological_process	lipid catabolic process	0.009193045
GO:0045017	biological_process	glycerolipid biosynthetic process	0.005724206
GO:0000271	biological_process	polysaccharide biosynthetic process	0.005799694
GO:0006094	biological_process	gluconeogenesis	0.002235584
GO:0046364	biological_process	monosaccharide biosynthetic process	0.007182709
GO:0006418	biological_process	tRNA aminoacylation for protein translation	0.005981961
GO:0030719	biological_process	P granule organization	0.008757257
GO:0006032	biological_process	chitin catabolic process	0.003842729
GO:0030030	biological_process	cell projection organization	0.00993608
GO:0044242	biological_process	cellular lipid catabolic process	0.008319245
GO:0034637	biological_process	cellular carbohydrate biosynthetic process	9.76E-06
GO:0071554	biological_process	cell wall organization or biogenesis	0.009837782
GO:0019319	biological_process	hexose biosynthetic process	7.15E-04
GO:0016051	biological_process	carbohydrate biosynthetic process	1.02E-04
GO:0015630	cellular_component	microtubule cytoskeleton	0.002655754
GO:0031224	cellular_component	intrinsic to membrane	6.88E-04
GO:0016020	cellular_component	membrane	0.006296849
GO:0015294	molecular_function	solute:cation symporter activity	0.002948052
GO:0005042	molecular_function	netrin receptor activity	0.008576329
GO:0004984	molecular_function	olfactory receptor activity	0.005925269

table S7 Comparison of evolutionary rates of *D. miranda* genes relative to *D. pseudoobscura*

	autosomes	chrX	neo-X	functional neo-Y	non-functional neo-Y
Studied gene pairs	6166	4312	2379	1103	902
K_a in %	1.8608(0.6804)	1.7138(0.6335)	1.2707(0.5995)	1.6030(1.0990)	2.4870(1.6380)
K_s in %	6.2842(4.8661)	5.6914(4.4451)	5.4247(4.6062)	5.8600(5.2760)	7.5480(6.4280)
K_a/K_s	0.2721(0.1512)	0.2835(0.1496)	0.2485(0.1345)	0.3092(0.2154)	0.3524(0.2697)
Fop	0.6300(0.6358)	0.6683(0.6748)	0.6504(0.6586)	0.6494(0.6556)	0.6410(0.6482)

This table shows the average (median) values of non-synonymous substitution rates (K_a), synonymous substitution rates (K_s) and frequency of optimal codons (Fop) for all the genes along each chromosome. Genes with pairwise K_a or K_s larger than 0.5 or K_a/K_s ratio higher than 5 were removed.

table S8 Sexually antagonistic genes on the neo-sex chromosomes of *D. miranda*. Significance was assessed using a series of Fisher’s exact tests by comparing the gene content of the ancestral neo-sex chromosome against either fast evolving neo-Y genes or non-functional neo-Y genes, to test for an enrichment of a specific fitness category among different gene sets (* *p*<0.05; ** *p*<0.01).

	investigated gene #	♂ fitness related	♀ fitness related	♂+♀-	♂-♀+
fast evolving neo-Y genes	312	22*	9	21*	8
non-functional neo-Y genes	1151	53	39	55	50**
ancestral neo-sex genes	2951	121	82	109	89

table S9 Enriched GO terms for genes with neo-X biased expression using whole adult male *D. miranda* transcriptome data

GO ID	Namespace	Name	P-value
GO:0006032	biological_process	chitin catabolic process	3.37E-05
GO:0044242	biological_process	cellular lipid catabolic process	1.01E-04
GO:0046165	biological_process	alcohol biosynthetic process	1.08E-04
GO:0016042	biological_process	lipid catabolic process	2.53E-04
GO:0034637	biological_process	cellular carbohydrate biosynthetic process	3.00E-04
GO:0042447	biological_process	hormone catabolic process	3.51E-04
GO:0019319	biological_process	hexose biosynthetic process	5.14E-04
GO:0006094	biological_process	gluconeogenesis	5.95E-04
GO:0006026	biological_process	aminoglycan catabolic process	5.96E-04
GO:0046364	biological_process	monosaccharide biosynthetic process	0.001352
GO:0016115	biological_process	terpenoid catabolic process	0.00202
GO:0016051	biological_process	carbohydrate biosynthetic process	0.002257
GO:0008300	biological_process	isoprenoid catabolic process	0.002948
GO:0051983	biological_process	regulation of chromosome segregation	0.003547
GO:0035233	biological_process	germ cell repulsion	0.004042
GO:0006716	biological_process	juvenile hormone metabolic process	0.004785
GO:0050919	biological_process	negative chemotaxis	0.005583
GO:0000272	biological_process	polysaccharide catabolic process	0.005859
GO:0015850	biological_process	organic alcohol transport	0.007161
GO:0005975	biological_process	carbohydrate metabolic process	0.007392
GO:0032008	biological_process	positive regulation of TOR signaling cascade	0.007742
GO:0005737	cellular_component	cytoplasm	0.002369
GO:0005739	cellular_component	mitochondrion	0.003738
GO:0044444	cellular_component	cytoplasmic part	0.004944
GO:0044422	cellular_component	organelle part	0.00905
GO:0003824	molecular_function	catalytic activity	1.66E-05
GO:0005044	molecular_function	scavenger receptor activity	0.00121
GO:0015665	molecular_function	alcohol transmembrane transporter activity	0.006737
GO:0016798	molecular_function	hydrolase activity, acting on glycosyl bonds	0.007026
GO:0008252	molecular_function	nucleotidase activity	0.00802
GO:0022892	molecular_function	substrate-specific transporter activity	0.009791

table S10 Enriched GO terms for genes with non-biased expression using whole adult male *D. miranda* transcriptome data

GO ID	Namespace	Name	P-value
GO:0007526	biological_process	larval somatic muscle development	0.009996
GO:0006807	biological_process	nitrogen compound metabolic process	0.006325
GO:0048869	biological_process	cellular developmental process	0.00257
GO:0042052	biological_process	rhabdomere development	0.003236
GO:0019222	biological_process	regulation of metabolic process	0.001836
GO:0023052	biological_process	signaling	0.004301
GO:0044260	biological_process	cellular macromolecule metabolic process	2.29E-06
GO:0050789	biological_process	regulation of biological process	1.96E-05
GO:0050794	biological_process	regulation of cellular process	0.002277
GO:0080090	biological_process	regulation of primary metabolic process	0.003821
GO:0051030	biological_process	snRNA transport	0.002899
GO:0065007	biological_process	biological regulation	9.82E-06
GO:0009892	biological_process	negative regulation of metabolic process	0.001328
GO:0007275	biological_process	multicellular organismal development	0.002672
GO:0006139	biological_process	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.69E-04
GO:0032501	biological_process	multicellular organismal process	4.54E-05
GO:0009987	biological_process	cellular process	3.11E-05
GO:0043170	biological_process	macromolecule metabolic process	0.00398
GO:0032502	biological_process	developmental process	1.11E-05
GO:0048519	biological_process	negative regulation of biological process	0.001082
GO:0050890	biological_process	cognition	0.006286
GO:0012505	cellular_component	endomembrane system	0.005167
GO:0000808	cellular_component	origin recognition complex	0.002323
GO:0005637	cellular_component	nuclear inner membrane	0.004013
GO:0046930	cellular_component	pore complex	0.007205
GO:0031300	cellular_component	intrinsic to organelle membrane	0.004952
GO:0005634	cellular_component	nucleus	0.003875
GO:0005664	cellular_component	nuclear origin of replication recognition complex	0.005367
GO:0022838	molecular_function	substrate-specific channel activity	0.003839
GO:0005217	molecular_function	intracellular ligand-gated ion channel activity	0.00549
GO:0043028	molecular_function	caspase regulator activity	0.003094
GO:0004497	molecular_function	monooxygenase activity	0.007181
GO:0008036	molecular_function	diuretic hormone receptor activity	0.004591
GO:0015081	molecular_function	sodium ion transmembrane transporter activity	0.003233
GO:0003676	molecular_function	nucleic acid binding	0.003049
GO:0001071	molecular_function	nucleic acid binding transcription factor activity	0.001484
GO:0005261	molecular_function	cation channel activity	0.002825
GO:0017056	molecular_function	structural constituent of nuclear pore	0.008358

GO:0022803	molecular_function	passive transmembrane transporter activity	0.002783
GO:0016740	molecular_function	transferase activity	0.004313
GO:0004948	molecular_function	calcitonin receptor activity	0.006221

table S11 Enriched GO terms for genes with neo-Y biased expression using whole adult male *D. miranda* transcriptome data

GO ID	Namespace	Name	P-value
GO:0009987	biological_process	cellular process	3.79E-05
GO:0000003	biological_process	reproduction	9.54E-05
GO:0007320	biological_process	insemination	1.45E-04
GO:0050789	biological_process	regulation of biological process	1.49E-04
GO:0032504	biological_process	multicellular organism reproduction	1.72E-04
GO:0022414	biological_process	reproductive process	2.82E-04
GO:0065007	biological_process	biological regulation	3.17E-04
GO:0051704	biological_process	multi-organism process	3.30E-04
GO:0048609	biological_process	reproductive process in a multicellular organism	6.10E-04
GO:0043226	cellular_component	organelle	7.40E-04
GO:0048519	biological_process	negative regulation of biological process	9.13E-04
GO:0009891	biological_process	positive regulation of biosynthetic process	0.001048829
GO:0031328	biological_process	positive regulation of cellular biosynthetic process	0.001246691
GO:0005488	molecular_function	binding	0.00163262
GO:0044087	biological_process	regulation of cellular component biogenesis	0.001795576
GO:0044422	cellular_component	organelle part	0.002164566
GO:0009893	biological_process	positive regulation of metabolic process	0.00231426
GO:0060537	biological_process	muscle tissue development	0.002370667
GO:0004222	molecular_function	metalloendopeptidase activity	0.002393485
GO:0043170	biological_process	macromolecule metabolic process	0.003116358
GO:0051173	biological_process	positive regulation of nitrogen compound metabolic process	0.003870398
GO:0032991	cellular_component	macromolecular complex	0.004382914
GO:0009953	biological_process	dorsal/ventral pattern formation	0.004435639
GO:0007620	biological_process	copulation	0.004441265
GO:0044424	cellular_component	intracellular part	0.004458212
GO:0044260	biological_process	cellular macromolecule metabolic process	0.004569214
GO:0016817	molecular_function	hydrolase activity, acting on acid anhydrides	0.005351765
GO:0003676	molecular_function	nucleic acid binding	0.006288673
GO:0005622	cellular_component	intracellular	0.006537098
GO:0051254	biological_process	positive regulation of RNA metabolic process	0.007389792
GO:0031325	biological_process	positive regulation of cellular metabolic process	0.007520772
GO:0007349	biological_process	cellularization	0.007714538
GO:0045935	biological_process	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.00794229
GO:0010604	biological_process	positive regulation of macromolecule metabolic process	0.009396109

table S12 Linear regression coefficients of absolute expression vs. *K*a/*K*s ratio

Linear Regression Coefficient	autosomes	chrXL/XR	hemizygous neo-X	diploid neo-X
Testis	0.1076***	0.1227***	0.1336***	0.09903***
Testis (exp. level>20)	0.1326***	0.2133***	0.2484***	0.1211**
Accessory Gland	0.03905***	-0.02081	0.06193*	0.00123
Acc . Gland (exp. level >20)	0.05152*	-0.04293	0.1653**	-0.03179
Male Carcass	-0.08869***	-0.1278***	-0.09674***	-0.11967***
Male Carcass (exp. level>20)	0.03970	0.02899	0.12476*	-0.01756
Ovary	-0.015043	-0.03914**	-3.333e-06	-0.06875**
Female Carcass	-0.043908***	-0.0698***	-0.06045*	-0.08923***

* reflect significance levels of linear regression: *** (*P*<0.001), ** (*P*<0.01), *(*P*<0.05), others (not significant).

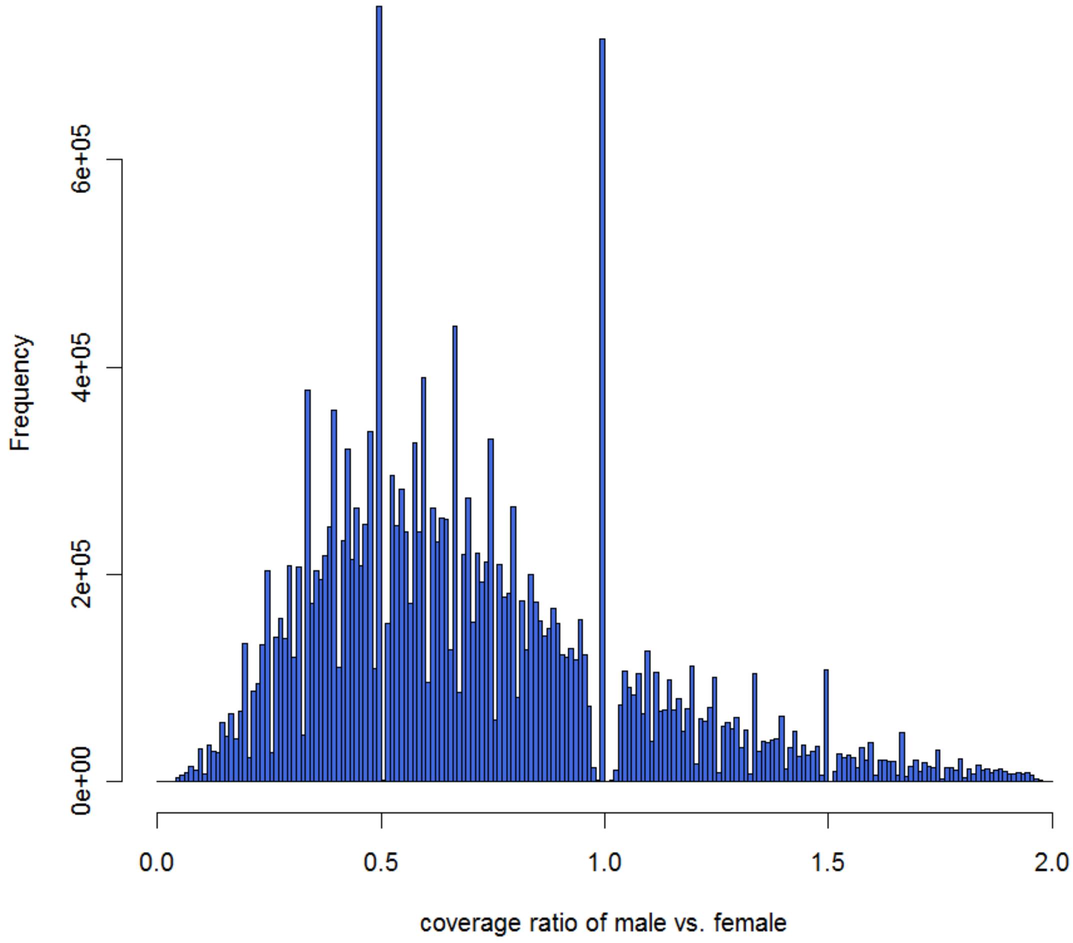

fig. S1. Frequency distribution of male vs. female coverage ratios.

We counted the number of reads mapped to each site of the neo-X scaffolds, and then investigated the frequency distribution of the ratio of male vs. female reads for all sites. The two pronounced peaks at 0.5 and 1 correspond to highly divergent and non-divergent sites between the neo-sex chromosomes.

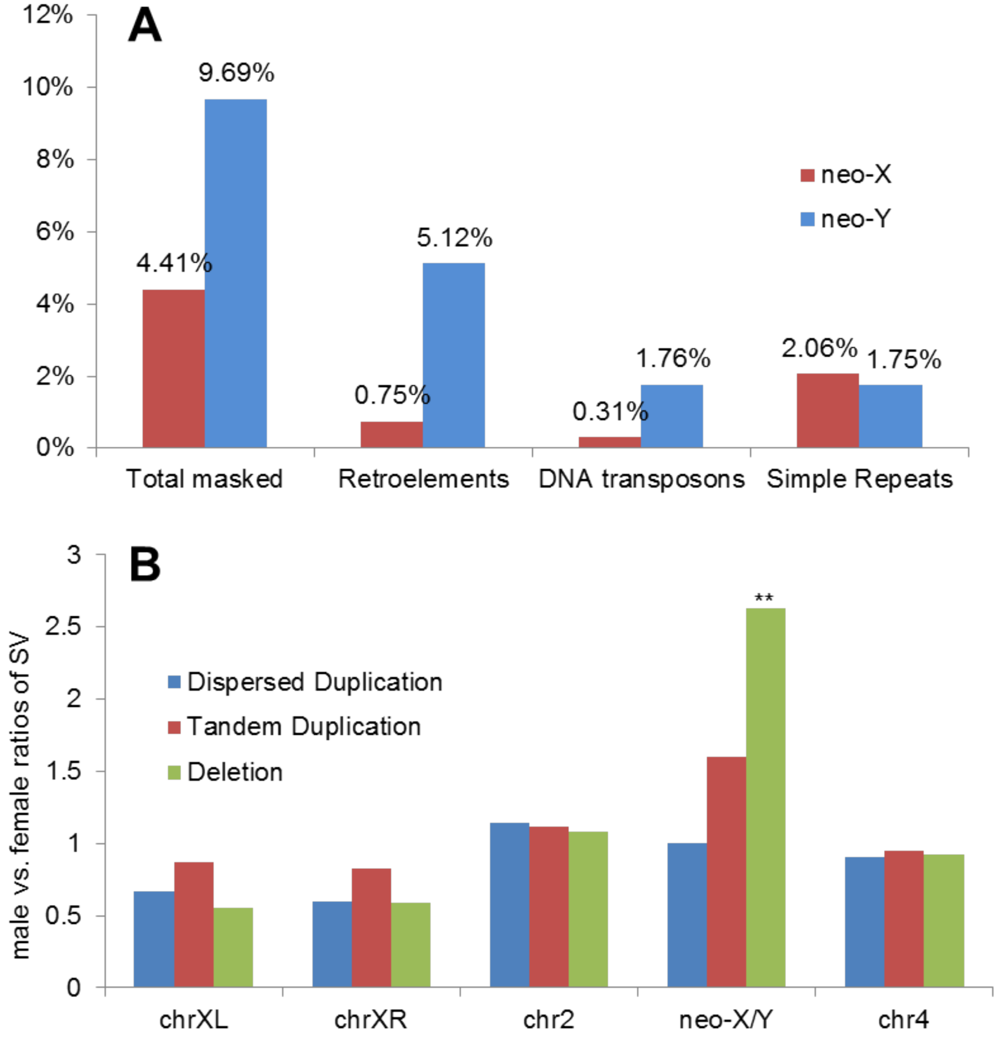

fig. S2. Genomic divergence between the neo-X and neo-Y

(A) We compared repeat contents of alignment gaps between homologous neo-X vs. neo-Y regions (i.e. insertions and deletions in either the neo-X or neo-Y chromosome). The neo-Y shows a 2-fold enrichment of repeats compared with the neo-X, which is mainly contributed by transposons or retroposons.

(B) Comparisons of structural variations derived from male data vs. female data. Their ratio along the neo-sex chromosomes should mainly reflect neo-X/Y divergence.

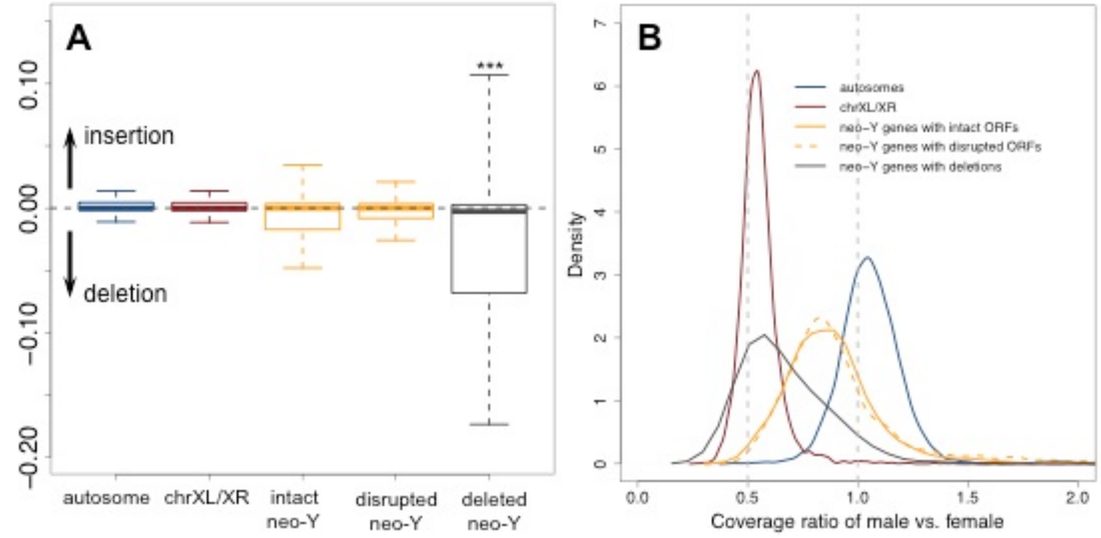

fig. S3. Validation of partly or completely deleted genes from the neo-Y. (A) Boxplots of length ratios of 454 male reads vs. their aligned regions in the Illumina assembly (excluding the neo-Y assembly). Neo-sex linked genes are divided into three categories: those with a functional neo-Y copy, a non-functional neo-Y copy, and those which are partly (or entirely) deleted from the neo-Y. Log-ratios lower than 0 indicate deletion events in the 454 reads, as expected if a partly or completely deleted neo-Y gene is aligned against its neo-X homolog. The length ratio of (partly) deleted neo-Y genes is significantly lower than that of other neo-sex genes (Wilcoxon test, P<0.001) **(B)** Density plots of read coverage ratios of male vs. female reads for different gene sets (using Illumina reads). Autosomes show equal read coverage between males and females while chrXL/XR show half of the female coverage in male. As expected, genes with neo-Y copies partially or completed deleted tend to have lower male vs. female ratio compared to other neo-sex linked genes.

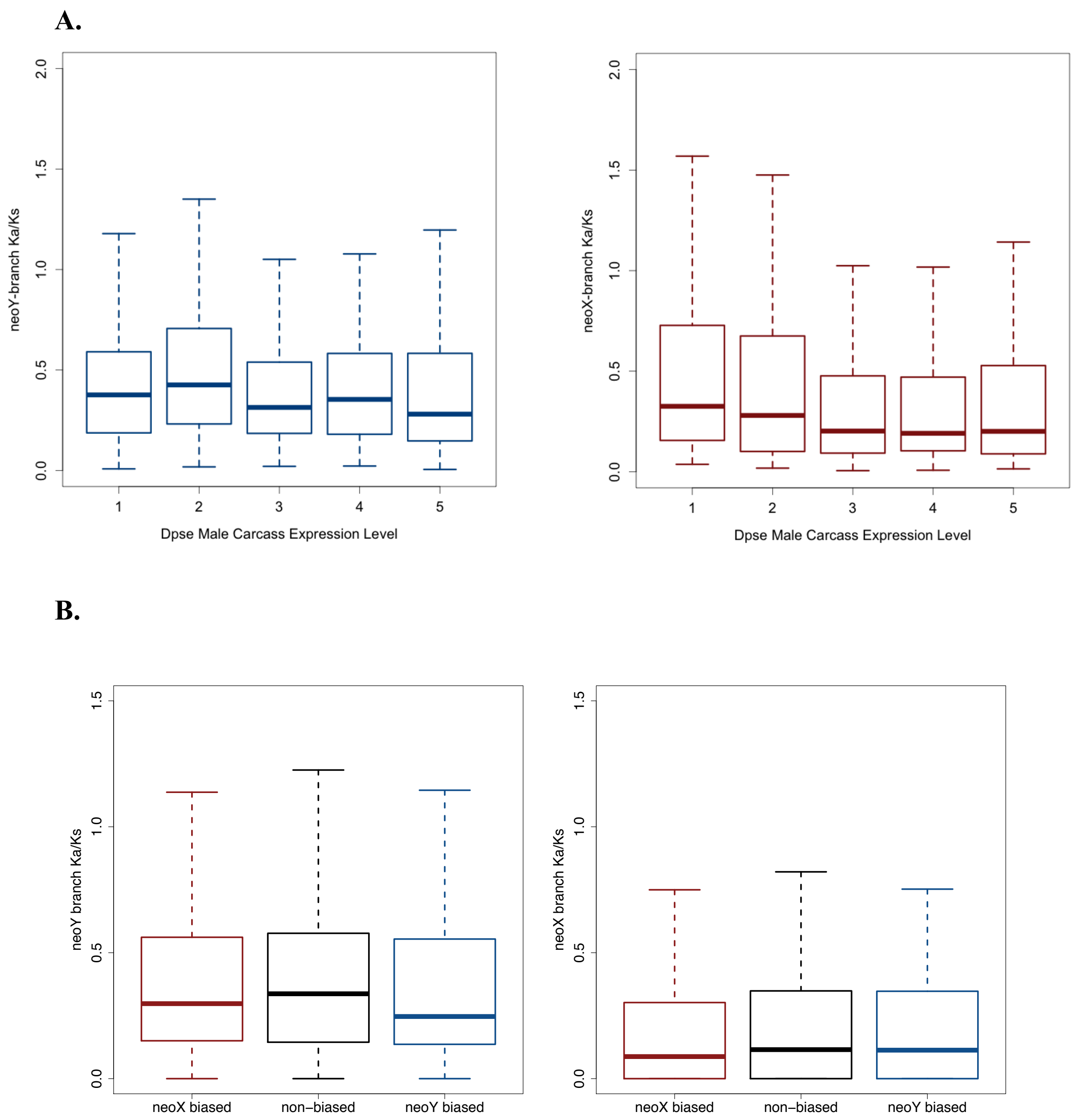

fig. S4. Protein divergence (K_a/K_s) along the neo-X and neo-Y chromosome for different expression categories. A. Neo-sex linked genes were grouped into five equal-sized bins according to ancestral expression using *D. pseudoobscura* male carcass data (1-low expression; 5-high expression). Neo-Y genes generally show increased levels of protein evolution relative to neo-X genes, independent of ancestral expression levels. **B**. Neo-sex linked genes were grouped into those showing neo-X, neo-Y or non-biased expression (shown is neo-sex biased expression assayed in accessory glands). Neo-Y genes generally show increased levels of protein evolution relative to neo-X genes, independent of expression bias.

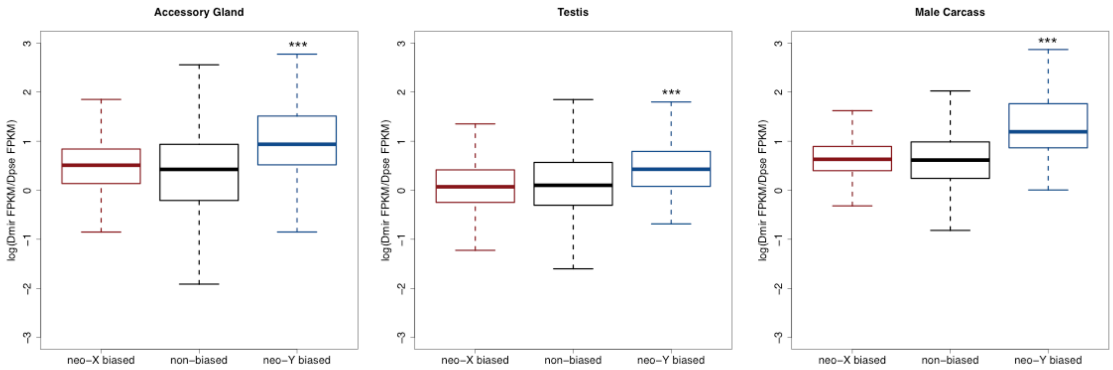

fig. S5 Relative expression of neo-X-biased, neo-Y-biased and non-biased genes between *D. miranda versus D. pseudoobscura* for different tissues. Genes with neo-Y biased expression are expressed at a significantly higher level in *D. miranda* relative to *D. pseudoobscura* relative to neo-X biased or non-biased genes, indicating that neo-Y biased expression is caused by transcriptional up-regulation of neo-Y genes in *D. miranda*.

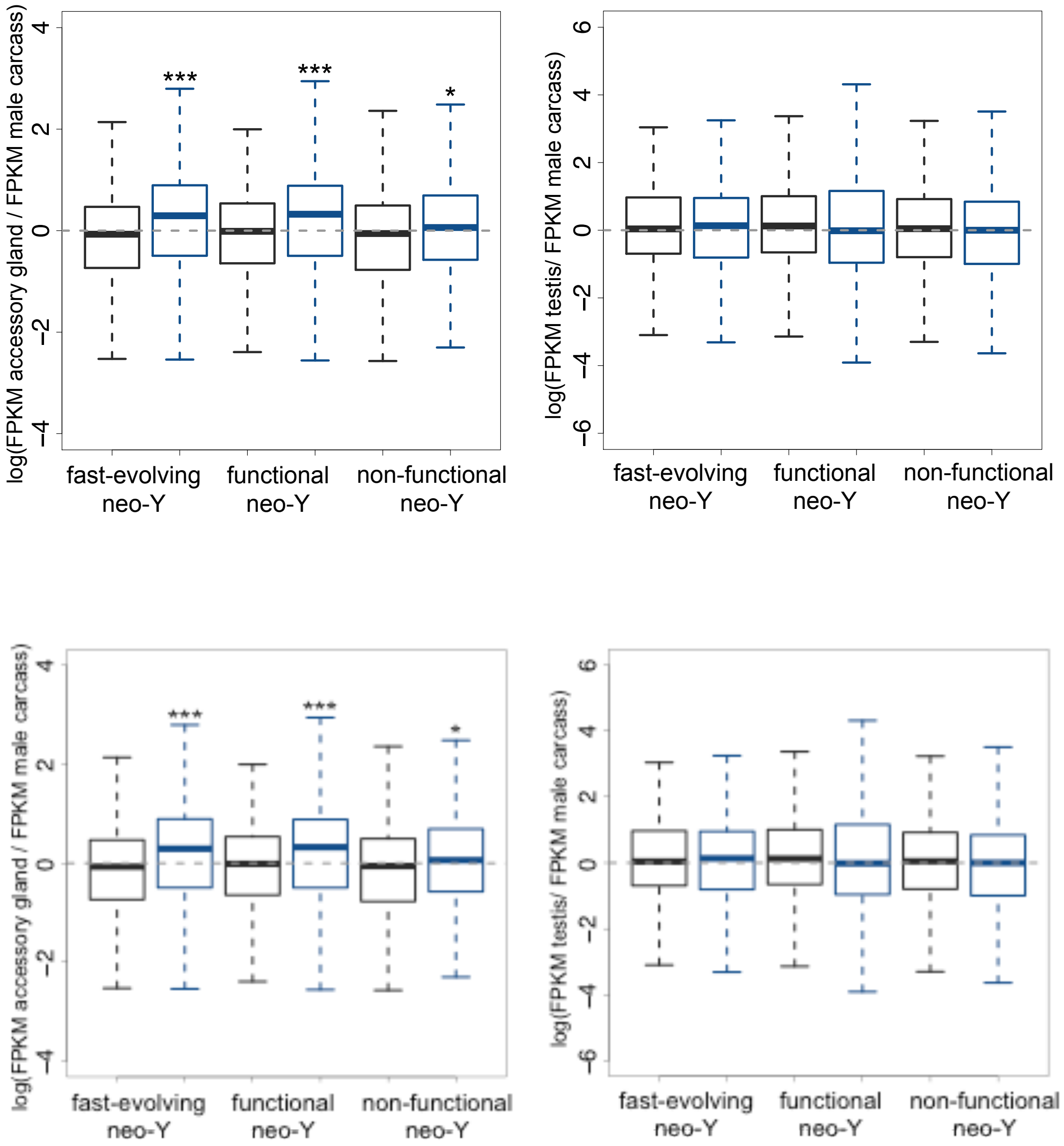

fig. S6 Expression specificity Tissue-specificity for testis and accessory gland genes (calculated as log(FPKM tissue / FPKM male somatic carcass)) of *D. miranda* genes (blue) compared to their *D. pseudoobscura* orthologs (black) for neo-Y genes (top) chrXL, chrXR and the autosomes (bottom). *P*-values of Wilcoxon tests between fast-evolving (0.0005596), functional (1.622e^{-06}) and non-functional (0.019) neo-Y genes versus their orthologous *D. pseudoobscura* genes are given.

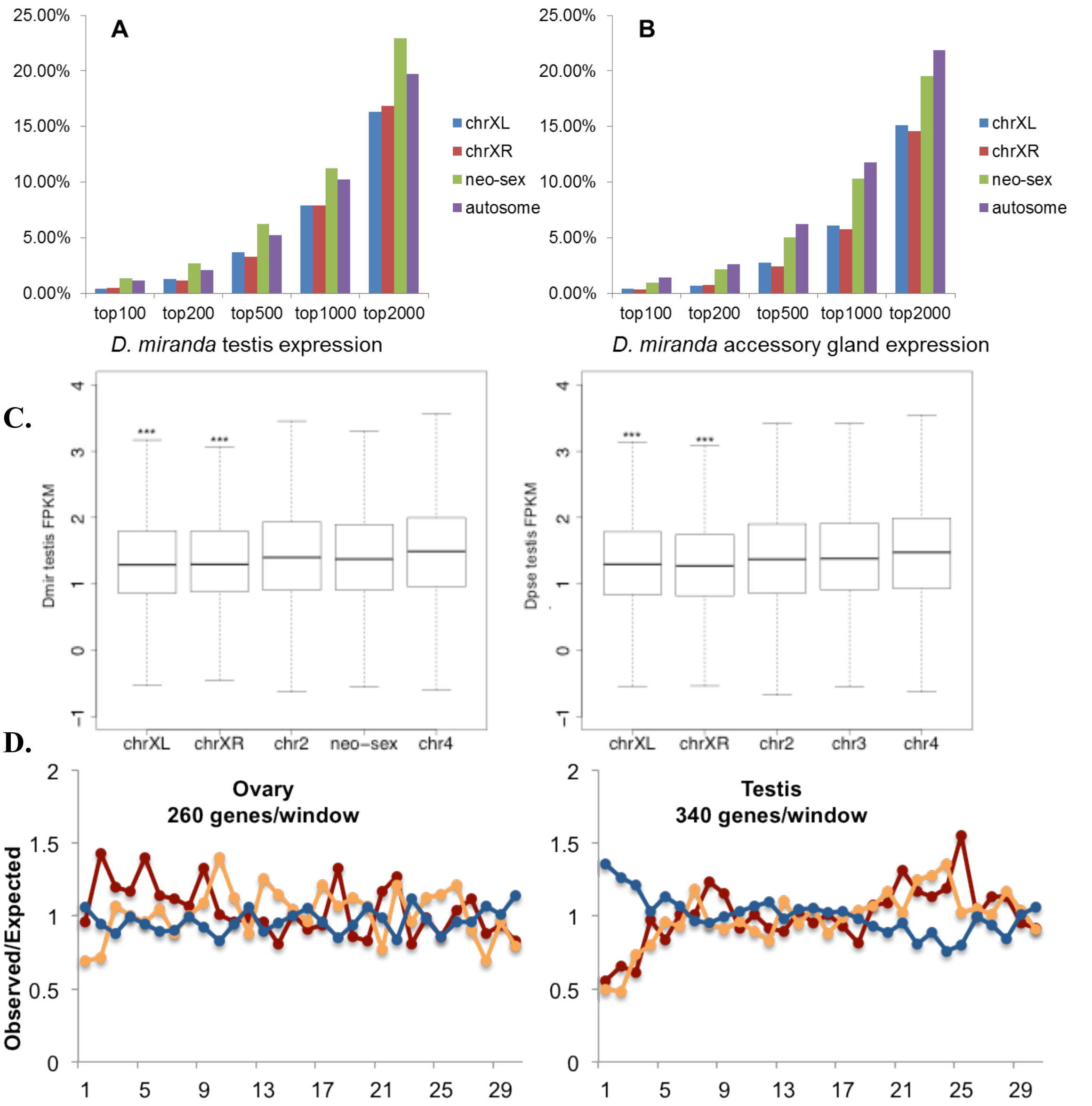

fig. S7 Demasculinization vs. feminization of X chromosomes. A. Chromosomal distribution of highly expressed testis and accessory gland genes in *D. miranda*. We investigated the proportions of highly expressed genes in testis and accessory glands from top 100 until top 2000 on each chromosome. ChrXL and chrXR consistently show a deficiency of genes highly expressed in testis and accessory glands compared to other chromosomes, regardless of the cut-off used. **C.** Log-based absolute expression levels (FPKM) from testis for each chromosome in *D. miranda* and *D. pseudoobscura*. **D.** Chromosomal distribution of ovary and testis genes binned by expression level in *D. miranda*. Genes were sorted according to their absolute expression level, and binned into 30 windows of equal size. For each window, the observed vs. expected number of genes in that window is calculated, assuming an even distribution. Blue lines are for autosomes, red lines for chrXL, and orange for chrXR.

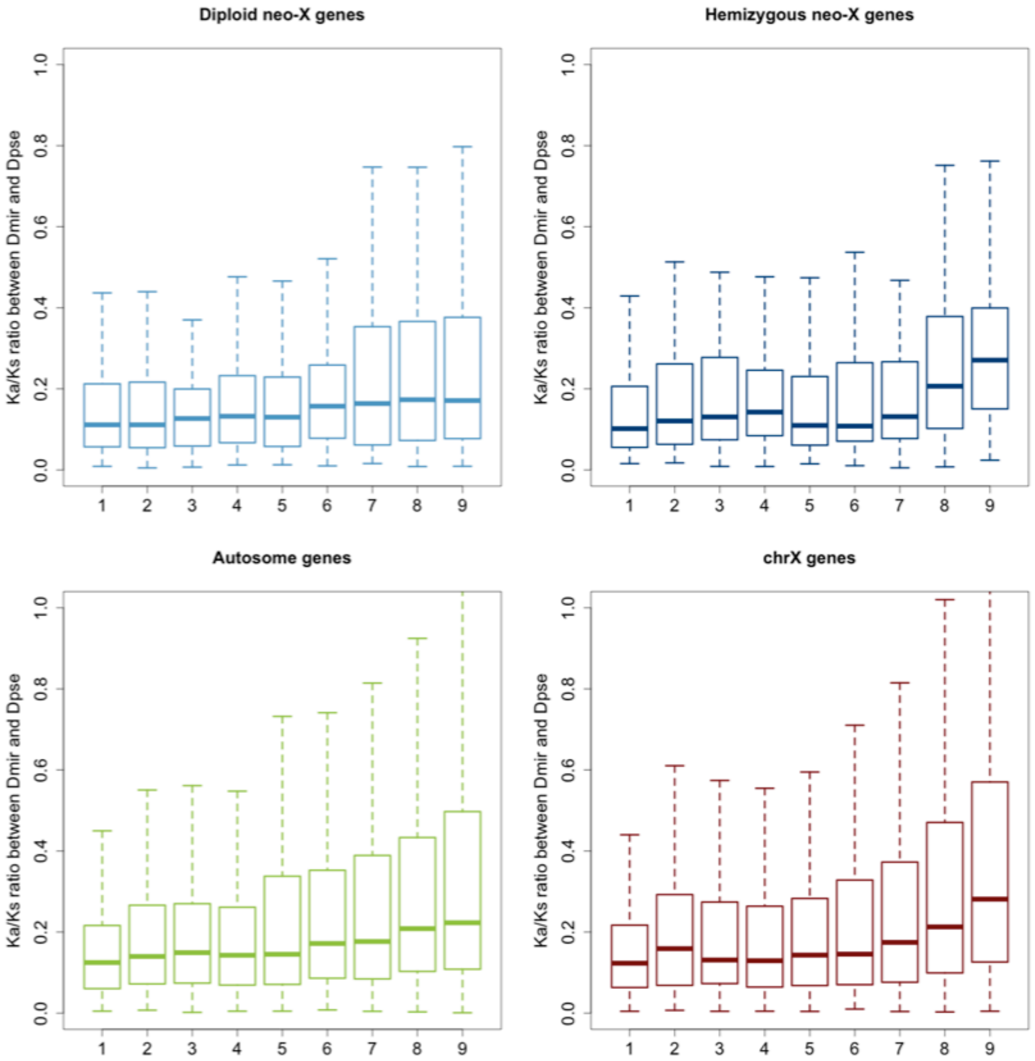

fig. S8 Correlation between testis expression levels and Ka/*K*s ratios.

Each box contains the same numbers of genes and expression level increases from left to right along the x-axis (1-lowly expressed; 10–highly expressed). The hemizygous neo-X genes are shown in grey.

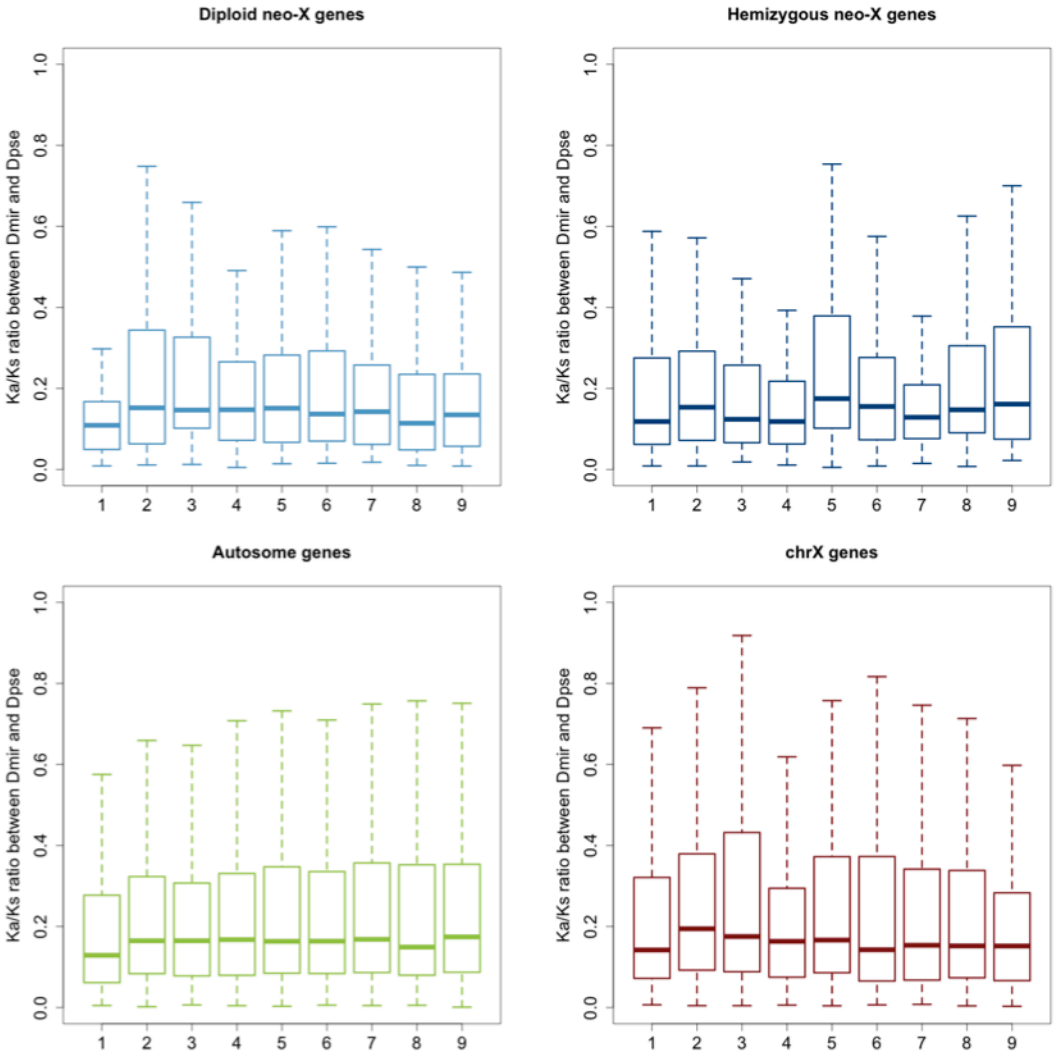

fig. S9 Correlation between accessory gland expression levels and *K*a/Ks ratios.

Each box contains the same numbers of genes and expression level increases from left to right along the x-axis (1-lowly expressed; 9–highly expressed).

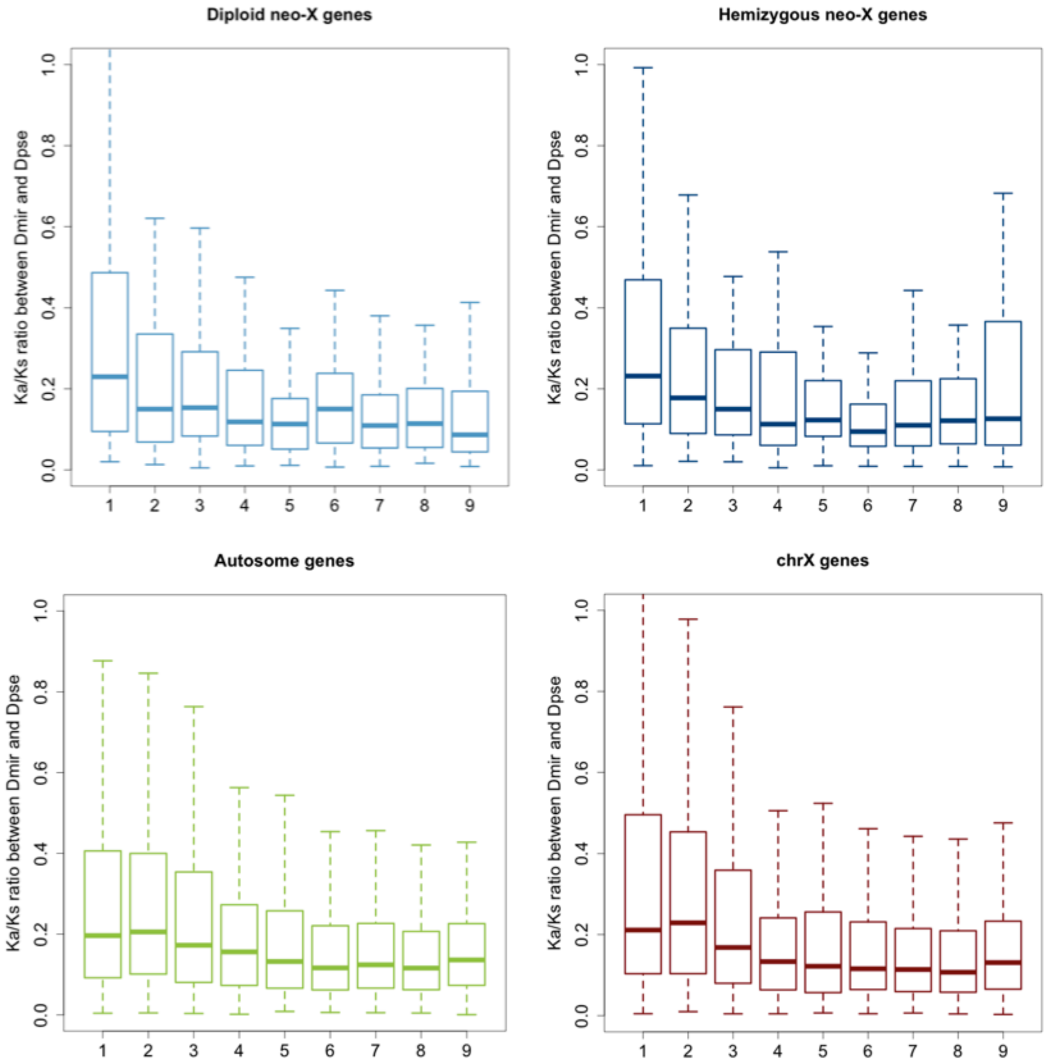

fig. S10 Correlation between male carcass expression levels and *K*a/*K*s ratios.

Each box contains the same numbers of genes and expression level increases from left to right along the x-axis (1-lowly expressed; 10–highly expressed). The hemizygous neo-X genes are shown in grey.

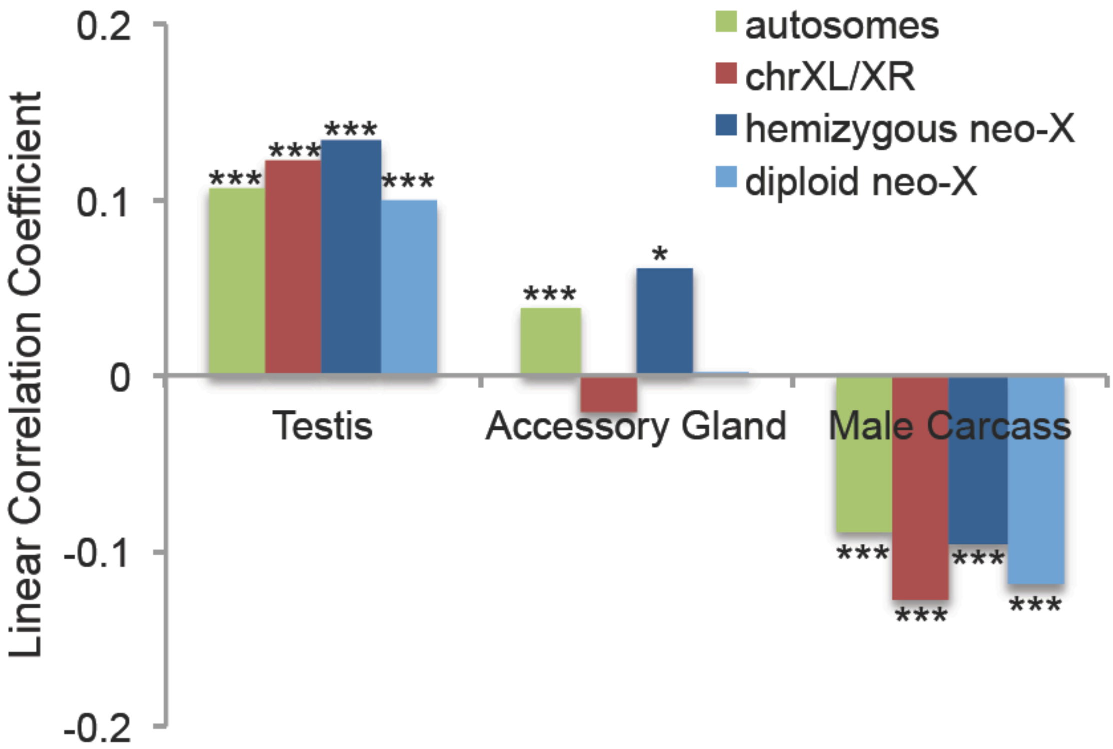

fig. S11 Linear correlation coefficients between tissue expression levels vs. *ω* ratio for sets of genes on different chromosomes and their significance levels. Hemizygous (but not diploid) neo-X genes show a highly significant positive correlation between their absolute expression levels in accessory gland and *ω* ratios (F-statistic comparing diploid vs. hemizygous neo-X genes, *P*<0.05). Genes more highly expressed in testis generally show faster rates of protein evolution, and this correlation is strongest for hemizygous neo-X genes (F-statistic comparing diploid vs. hemizygous neo-X genes, *P*=0.076). On the contrary, all chromosomes exhibit a significant negative correlation in male somatic carcass and female tissues ($P<10^{-3}$), indicating purifying selection on ubiquitously highly expressed genes (*43*). Masculinization of young X genes can be detected more readily on the neo-X than the ancestral X because the burst of adaptive evolution tends to be recent (*44*). Autosomal genes show a significant positive correlation between *ω* and expression levels in both testis and accessory glands, suggesting that diploidy itself is not sufficient to prevent adaptation at male-genes (that is, a significant fraction of male advantageous alleles are dominant). In contrast to autosomes, diploid neo-X genes show female-biased transmission, which may oppose selection for male-beneficial alleles at such genes. '*' (*P*<0.05), '**' (*P*<0.01), '***' (*P*<0.0001).

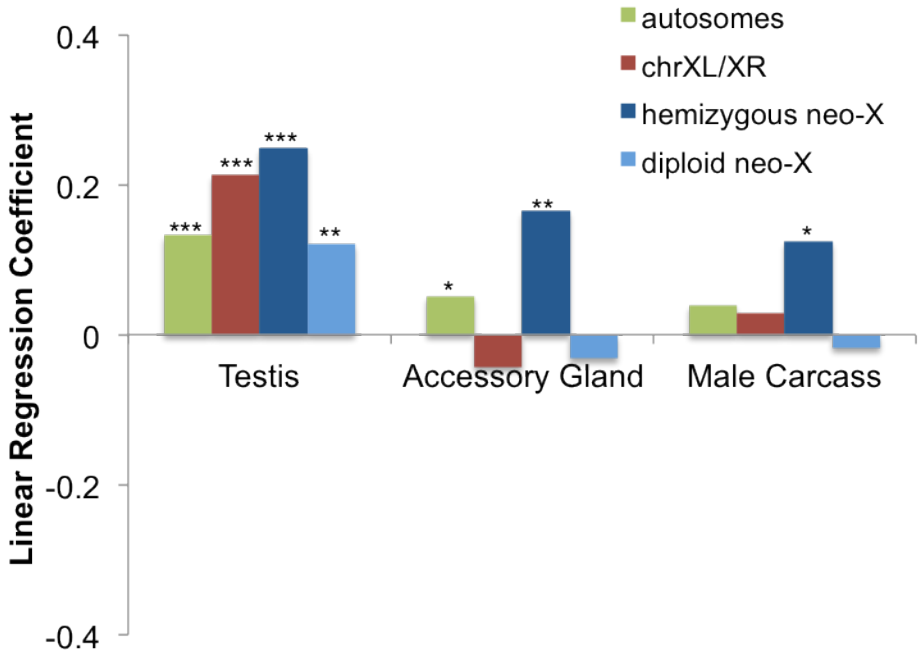

fig. S12 Linear regression of tissue expression levels vs. *ω* ratio for highly expressed genes. Highly expressed genes are defined as those with FPKM > 20.

References

26. R. Li *et al.*, *Genome Res.* **20**, 265 (Feb, 2010).
27. A. Bhutkar *et al.*, *Genetics* **179**, 1657 (Jul, 2008).
28. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Journal of molecular biology* **215**, 403 (Oct 5, 1990).
29. R. Li, Y. Li, K. Kristiansen, J. Wang, *Bioinformatics* **24**, 713 (2008).
30. R. Li *et al.*, *Genome Res* **19**, 1124 (Jun, 2009).
31. S. Marion de Procé, D. Halligan, P. Keightley, B. Charlesworth, *Journal of Molecular Evolution* **69**, 601 (2009).
32. E. Birney, M. Clamp, R. Durbin, *Genome research* **14**, 988 (May, 2004).
33. M. Krzywinski *et al.*, *Genome Res* **19**, 1639 (Sep, 2009).
34. J. M. Cridland, K. R. Thornton, *Genome Biol Evol* **2**, 83 (2010).
35. S. Bauer, S. Grossmann, M. Vingron, P. N. Robinson, *Bioinformatics* **24**, 1650 (Jul 15, 2008).
36. F. Abascal, R. Zardoya, M. J. Telford, *Nucleic acids research* **38**, W7 (Jul, 2010).
37. Z. Yang, R. Nielsen, *Mol Biol Evol* **17**, 32 (Jan, 2000).
38. Z. Yang, R. Nielsen, *Mol Biol Evol* **19**, 908 (Jun, 2002).

39. Z. Yang, *Mol Biol Evol* **15**, 568 (May, 1998).
40. S. Vicario, E.N. Moriyama, J.R. Powell, *BMC Evol Bio* **7**, 1471 (November 2007)

41. C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **25**, 1105 (May 1, 2009).
42. C. Trapnell *et al.*, *Nat Biotechnol* **28**, 511 (May, 2010).
43. A. M. Larracuente *et al.*, *Trends Genet.* **24**, 114 (Mar, 2008).

44. D. Bachtrog, J. D. Jensen, Z. Zhang, *PLoS Biol.* **7**, e82 (Apr 14, 2009).