Enrichment of hard sweeps on the X chromosome compared to autosomes in six *Drosophila* species

Mariana Harris[1], Bernard Kim[2], and Nandita Garud[3,4,*]

[1] Department of Computational Medicine, University of California Los Angeles, Los Angeles California, United States of America
[2] Department of Biology, Stanford University, Stanford, California, United States of America
[3] Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles California, United States of America
[4] Department of Human Genetics, University of California, Los Angeles, California, United States of America

*Correspondence to ngarud@ucla.edu

Abstract

The X chromosome, being hemizygous in males, is exposed one third of the time increasing the visibility of new mutations to natural selection, potentially leading to different evolutionary dynamics than autosomes. Recently, we found an enrichment of hard selective sweeps over soft selective sweeps on the X chromosome relative to the autosomes in a North American population of *Drosophila melanogaster*. To understand whether this enrichment is a universal feature of evolution on the X chromosome, we analyze diversity patterns across six commonly studied *Drosophila* species. We find an increased proportion of regions with steep reductions in diversity on the X chromosome compared to autosomes, accompanied by elevated homozygosity and an increased proportion of singletons. To assess if these signatures are consistent with positive selection, we simulate a wide variety of evolutionary scenarios spanning variations in demography, mutation rate, recombination rate, background selection, hard sweeps, and soft sweeps, and find that the diversity patterns observed on the X are most consistent with hard sweeps. Our findings highlight the importance of sex chromosomes in driving evolutionary processes and suggest that hard sweeps have played a significant role in shaping diversity patterns on the X chromosome across multiple *Drosophila* species.

Significance

Understanding the unique evolutionary dynamics of the X chromosome is a longstanding question in evolutionary biology given its central role in speciation, fertility, and sexual dimorphism. Recently, we found that hard sweeps, which are indicative of gradual adaptation, are more common on the X chromosome than on the autosomes in a North American population of *Drosophila melanogaster*. This is because new mutations on the X chromosome are more visible to natural selection due to male hemizygosity. Here, we analyze multiple whole genome sequences from six *Drosophila* species and find that this pattern is common across all species. Our study demonstrates the power of comparative population genetics in unraveling the forces shaping genetic variation within and between species.

Introduction

The X chromosome has long been a subject of substantial interest in evolutionary biology due to its unique features that set it apart from the autosomes. Notably, the X harbors genes responsible for speciation, fertility, sexual dimorphism, and brain function (1–6), highlighting its biological importance. Moreover, previous work suggests that the X chromosome may serve as a potential target of sexually antagonistic selection (7–9), further emphasizing its significance in evolution. Consequently, studying adaptation on the X chromosome, and how it differs from that of autosomes, can provide insights into the mechanisms driving genetic diversity, sexual selection, and speciation, thereby deepening our understanding of the broader processes that shape genetic variation across populations.

Adaptation on the X may differ from that of the autosomes due to two key differences. First, the X is expected to have a lower effective population size (N_{eX}) compared to autosomes (N_{eA}), leading to a decreased influx of new mutations. Second, due to male hemizygosity, new mutations on the X of males are immediately exposed to natural selection. This increased exposure to selection may lead to a higher probability of fixation of new recessive beneficial mutations ("Faster-X" effect (10)) and a more efficient purging of deleterious variation on the X compared to autosomes, leading to lower levels of standing genetic variation on the X. Thus, as a consequence of these two factors, at the onset of positive selection, there will be a lower adaptive mutational supply on the X, resulting in more gradual rates of adaptation, or in other words, fewer haplotypes rising to high frequency bearing the adaptive allele (11–14).

Gradual adaptation leaves behind signatures known as hard selective sweeps, whereby a single adaptive mutation rises in frequency, resulting in deep dips in diversity in the vicinity of the adaptive locus (15, 16). In contrast, rapid adaptation is characterized by soft selective sweeps, where multiple adaptive mutations on distinct haplotypes sweep through the population simultaneously, not necessarily causing dips in diversity (17–20). Recently, we found evidence of an enrichment of hard sweeps on the X chromosome compared to autosomes in a North American population of *D. melanogaster* (21), suggesting that the X chromosome is subject to different evolutionary dynamics than the autosomes. Whether enrichment of hard sweeps on the X is a universal feature of molecular evolution has yet to be determined, as only a few species have been shown to have a higher prevalence of hard sweeps on the X compared to the autosomes (21, 22). Quantifying the prevalence of hard versus soft sweeps in natural populations has been of great

interest and debate (23–29). Thus, understanding the prevalence of hard and soft sweeps more broadly is crucial as it can shed light on common mechanisms that underlie adaptation in natural populations.

To understand if the X is generically enriched for hard sweeps in many species, we analyze population genomic data from six *Drosophila* species. By leveraging whole genome samples from multiple species and populations (22, 30, 31), we can identify trends that are the norm across species, as well as exceptions that are indicative of the unique biology of individual species. However, four of the populations we analyze in this study have small sample sizes (*n*=7-23 samples), making it difficult to conduct the same haplotype-based scan (21, 32) used in our previous work (21). Additionally, each of these species has a unique demographic history for which we do not have accurate models, rendering our previous simulation-based approach for classifying putative sweeps as hard and soft challenging (21). To overcome these challenges, we use a combination of single-nucleotide diversity and haplotype homozygosity statistics to analyze the patterns of diversity on the autosomes and the X chromosome across species, specifically looking for evidence of hard sweeps on the X that is inconsistent with other neutral and selective forces including background selection (BGS) and soft selective sweeps. Our empirical and simulation analyses show evidence that hard sweeps have played a significant role in shaping diversity patterns on the X chromosome in multiple *Drosophila* species suggesting that hard sweeps on the X are the norm rather than the exception.

Results

We analyzed population genomic data of six different *Drosophila* species from seven populations, including: *D. melanogaster* (n=100 Zambia (ZI), n=100 Raleigh (RA)), *D. simulans* (n=170), *D. sechellia* (n=23), *D. mauritiana* (n=15), *D. santomea* (n=7) and *D. teissieri* (n=11) (**Table S1**). We include the presumed ancestral population from Zambia as well as the derived population from Raleigh from *D. melanogaster* to understand not only how selection varies across species, but also across populations within a species. First, we compare the patterns of nucleotide diversity (π) between the autosomes and the X chromosome in the data. Next, we analyze the proportion of singletons and haplotype homozygosity in low diversity regions on the X compared

to the rest of the chromosome. Finally, we simulate a variety of models, both without and with selection, to test whether any of these can generate the patterns observed in the data.

Patterns of X versus autosomal diversity across the genomes of six *Drosophila* species

When a population is not subject to any selective forces, has the same mutation rate across sexes, a male-to-female ratio of one, and a constant effective population size (N_e), the diversity on the X chromosome (π_X) is expected to be 3/4 of the autosomal diversity (π_A) (11). However, such a simple model rarely captures the complex dynamics of natural populations, rather, it is likely that different evolutionary forces are at play, leading to deviations from the π_X/π_A=0.75 expectation. Across the seven populations studied, we observed a significantly lower diversity on the X compared to autosomes with five out of seven populations showing π_X/π_A below 0.75 (**Fig. 1**). The other two populations (*D. melanogaster* (ZI) and *D. teissieri*) also deviated from the π_X/π_A=0.75 expectation but in the opposite direction, with π_X/π_A>0.75. This trend has been previously reported for African populations of *D. melanogaster* and has been attributed to other evolutionary forces such as female sex bias (33–35).

Functionally important regions of the genome, such as exons, are expected to be subject to selection, positive or negative (22, 30, 36). Hence, to investigate the influence of selection across species, we divided the data into three regions: exons, introns, and intergenic regions and computed the ratio of X to autosomal diversity (π_X/π_A) for each region. We found a stronger reduction in diversity in exons compared to introns and intergenic regions for most species (**Figure 1**)**,** pointing to the functional importance of exons (22). Moreover, we found that diversity was consistently lower on the X compared to autosomes for all regions for most species (**Figure 1**), suggesting more directional selection on the X compared to autosomes. However, we noted an exception in *D. sechellia,* where exons exhibited higher levels of diversity than introns and intergenic regions. This behavior could be explained by the high levels of introgression from *D. simulans* to *D. sechellia* (37–39). Introgression tends to be higher in non-functional regions of the genome and is generally higher on the autosomes than on the X due to the involvement of sex chromosomes in hybrid incompatibilities (40–42). This could lead to a greater difference between X and autosomal diversity in intergenic regions compared to exons, potentially explaining the patterns observed in *D. sechellia*'s data.

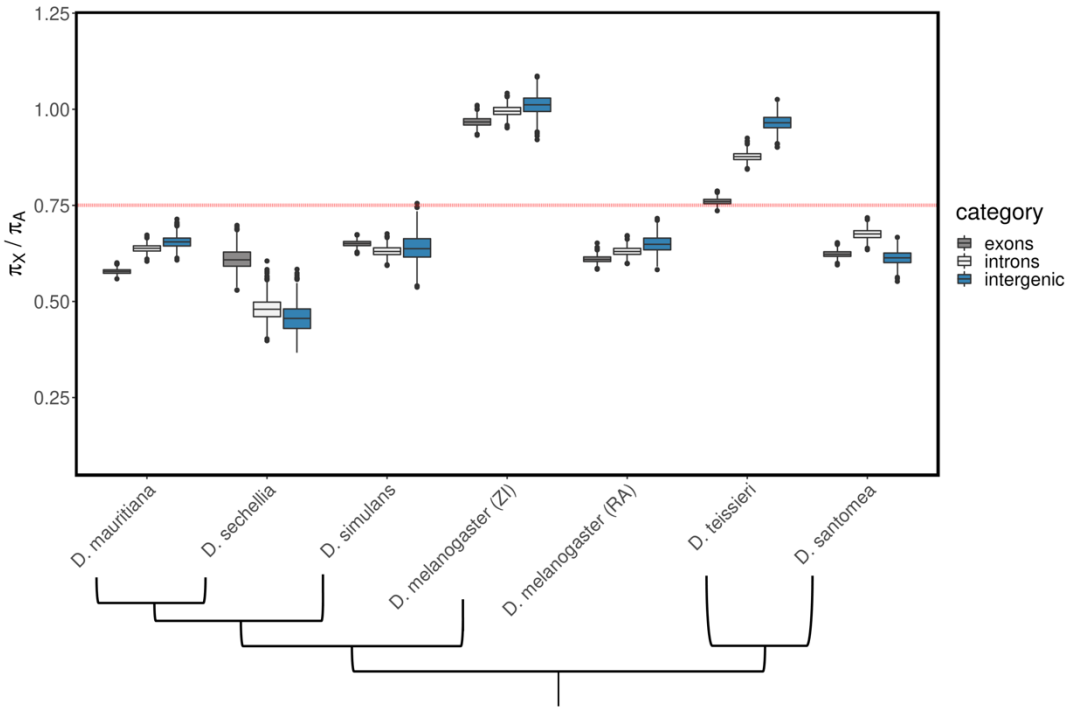

Figure 1. Nucleotide diversity (π/bp) in exons, introns and intergenic regions on the autosomes and the X chromosome of six *Drosophila* species. π_X/π_A is lower on exons than on introns and intergenic regions in most species. The red line corresponds to the 0.75 neutral expectation. For each category we performed 1,000 bootstrap replicates (**Methods**). The tree shown below the X-axis represents the phylogenetic relationship between species as described in FlyBase (43).

To identify potential regions under selection, we calculated π/bp in 20Kb windows across each chromosome for every species. As expected, we observed a range of diversity values across the X chromosome of each species. Windows with the lowest levels of diversity have diversity levels between 0.5% to 10% of the mean diversity on the X of the respective species (**Table S2**). With this range in mind, windows were labeled as low diversity windows if they fell below a diversity threshold set as 20% of the X chromosomal π/bp average. Importantly, we computed this threshold for each chromosome separately to normalize for variance in coalescence time across chromosomes. We found that, except for *D. teissieri*, the X chromosome showed a higher

proportion of windows below the low diversity threshold compared to the autosomes (**Figures 2A**, **2B** and **S1**). Remarkably, these low diversity windows often exhibited π/bp values significantly lower than the defined threshold with values as low 0.5% of the chromosomal average (**Table S2; Fig. S2**). Furthermore, we observed multiple instances of consecutive low diversity windows extending up to 200Kb, shown as the yellow shaded regions of **Figure 2A**. To validate our findings, we repeated this analysis using 50Kb and 100Kb analysis windows as well as two different low diversity thresholds. In all cases, we observed an increased proportion of low diversity windows on the X compared to autosomes (**Figure S3**).

If selection is indeed responsible for the excess of low diversity windows on the X, then the site frequency spectrum (SFS) in low diversity regions should show an enrichment of singletons (44–47). To test this, we computed the number of singletons on the X chromosome in windows below and above the low diversity threshold. We found a higher proportion of singletons in low diversity windows on the X (**Figure 2C**), consistent with the expected SFS under purifying or positive selection.

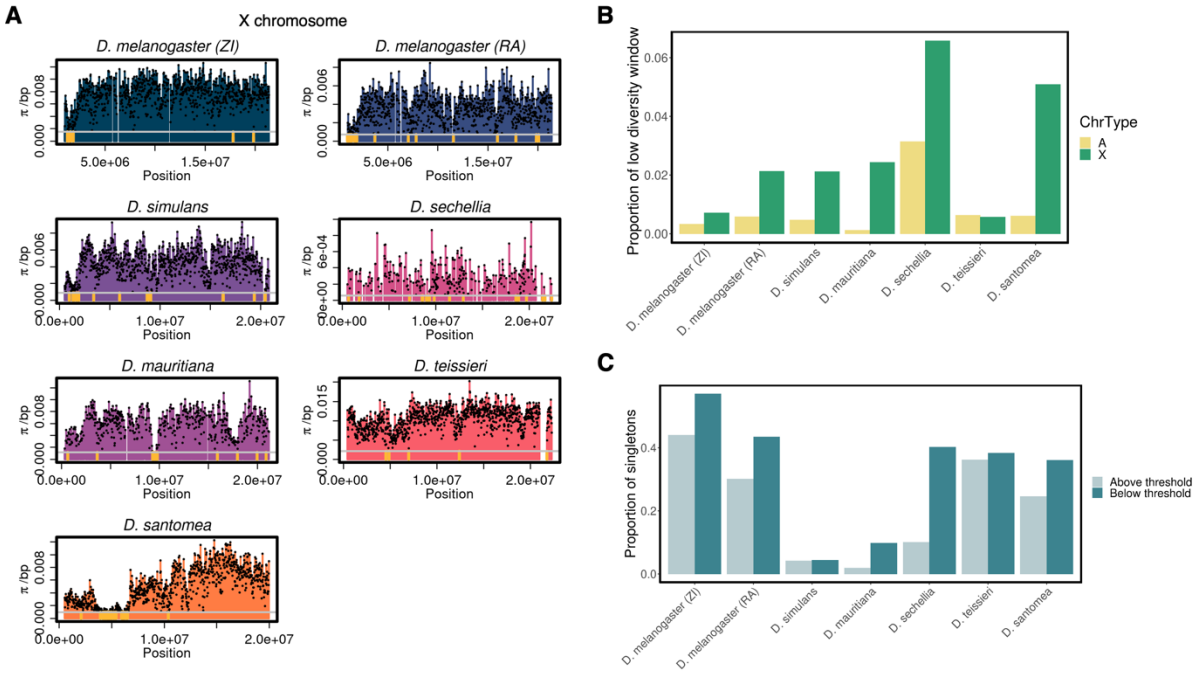

Figure 2. Multiple *Drosophila* species have a higher proportion of windows on the X with steep depressions of diversity compared to autosomes. (A) π/bp in 20kb windows on the X

chromosome of six *Drosophila* species and two *D. melanogaster* populations from Raleigh (RA) and Zambia (ZI). Windows where diversity is below 20% of the chromosomal average (grey horizontal line) are highlighted in gold. In **Figure S1**, the same figure for all autosomes is plotted for each species. **(B)** Proportion of windows below the 20% of the chromosomal average on the autosomes (yellow) and the X chromosome (green). **(C)** Proportion of singletons in windows below (light blue) and above (dark blue) the low diversity threshold defined as 20% of the chromosomal average on the X chromosome.

Elevated homozygosity can arise from different processes such as positive selection or population bottlenecks (48). Therefore, to gain deeper insights into the potential mechanisms responsible for the dips in diversity on the X, we computed haplotype homozygosity (H) across windows below and above the low diversity threshold (**Methods**; **Fig. 3**) in species with phased genomes (*D. melanogaster, D. simulans* and *D. mauritiana*). For species with unphased genomes, we computed multilocus genotype identity (G) ((49); **Methods**). We found that in all populations, windows below the low diversity threshold showed a significantly elevated homozygosity compared to the rest of the chromosome (one-sided Wilcoxon rank sum test p-val< 0.001; **Fig. 3**). Below we examine the role of selection in generating these differential patterns between the X and the autosomes.

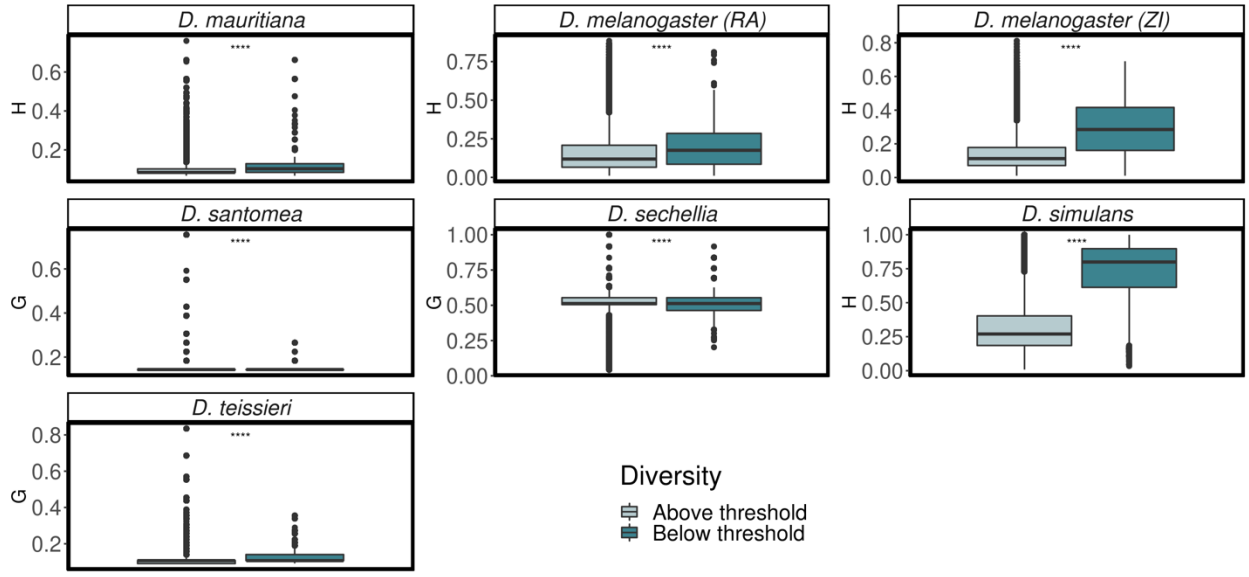

Figure 3. Homozygosity in windows below and above 20% of the average π/bp on the X

chromosome. Windows with π/bp below 20% of the chromosomal mean show elevated haplotype homozygosity (H) or elevated multilocus genotype homozygosity (G) compared to windows with π/bp above this threshold. Low diversity windows show a significant elevation in homozygosity compared to the rest of the X chromosome across all species (one-sided Wilcoxon rank sum test P-val< 0.001).

Neutral models cannot recapitulate diversity patterns observed in the data

The differing diversity patterns on the X versus autosomes could be the result of distinct population genetic forces including: (1) differences in mutation rates between males and females, (2) sex biased demography, (3) demographic forces differentially affecting X and autosomal diversity due to a stronger effect of drift on the X given its smaller N_e and/or (4) more efficient natural selection on the X due to male hemizygosity (13, 22, 30, 50).

To gain a better understanding of how different evolutionary processes influence the relative X to autosomal diversity, we quantified π_X, π_A and haplotype homozygosity across a wide variety of scenarios simulated with the population genetics simulator SLiM (51). We started by simulating neutral models of sex bias, mutation bias and population bottlenecks.

From these scenarios (**Fig. 4A**), we found that female bias can increase π_X/π_A above the 0.75 expectation, recapitulating the diversity observed in the Zambian population of *D. melanogaster* and consistent with evidence of female bias in African *D. melanogaster* populations previously reported in the literature (33–35, 52, 53) **(Fig. 1** and **Fig 4A)**. *D. teissieri* also has π_X/π_A>0.75, which could also be consistent with female bias. However, none of our female bias models produced elevated haplotype homozygosity compared to neutrality (**Fig. 4B**).

Additionally, we found that male sex bias can decrease π_X such that $\pi_X/\pi_A < 0.75$, which could potentially explain low π_X values observed in some species (*D. melanogaster (RAL), D. simulans, D. mauritiana* and *D. santomea;* **Fig. 1**). However, only the most extreme case of male bias (1:7 female-male bias) can marginally increase haplotype homozygosity compared to neutrality (**Fig. 4B**). To the best of our knowledge, such an extreme male-sex bias in the species studied has not been reported in the literature and therefore seems unlikely.

Lower X-linked mutation rates could result in stronger dips in diversity on the X compared to autosomes. A lower X-linked mutation rate is expected when there is a higher number of germline

cell divisions in males than females (54, 55). However, in *Drosophila,* the number of cell divisions in the male and female germlines has been shown to be similar (54), suggesting that there should be no significant difference in the mutation rate between the sexes. Nevertheless, other factors may differentially affect X-linked and autosomal mutation rates such as selection on codon usage on the X, given a higher level of codon usage bias and GC content on the X across many *Drosophila* species (56–62). Past studies on *D. melanogaster* have not shown a statistically significant difference between X-linked and autosomal mutation rates (58, 60, 63). For other species, evidence of a lower X-linked mutation rate is inconclusive, with few species showing evidence of lower X mutation rates (64), however more studies are needed to better understand X-linked versus autosomal mutation rate differences.

Nonetheless, we performed simulations to test the impact of a lower mutation rate on the X and whether this could, on its own, explain the patterns observed in the data. Simulations show that only the most extreme case of lower X-linked mutation rate (μ_X=0.1μ_A) can decrease diversity below the 20% of $\pi_{neutral}$, where $\pi_{neutral}$ is the average π/bp from neutral simulations. However, none of our low X-linked mutation rate simulations significantly elevated haplotype homozygosity. In fact, the only case that can considerably reduce π_X, μ_X =0.1μ_A, shows the lowest haplotype homozygosity across all models (**Fig. 4C**). From the above, we conclude that it is unlikely that the low diversity windows on the X are uniquely explained by a lower X-linked mutation rate.

Next, we tested the effects of demography on X versus autosome diversity. We tested two variations of a bottleneck model: a severe and short bottleneck and a shallow and long bottleneck (**Methods**). Both models were fitted to π/bp and S/bp in short introns from the Drosophila Genetic Reference panel (DGRP) dataset (24, 32). Our simulations revealed a π_X/π_A slightly below 0.75 as well as a significant increase in haplotype homozygosity in both bottleneck scenarios (**Fig 4A,B**). To test whether the higher variance on the X, expected from its smaller N_e, could result in local strong dips in diversity we looked at whether the distribution of π_X could achieve values below 20% of the average π_X for each model. In **Fig. 4D** we see that the 20% of the average π_X is well below the tail of the π_X distribution, suggesting that bottlenecks are unlikely to generate the strong local dips in π that we observe in the data (**Fig 2, Fig. S2**).

Finally, inbreeding is another process that could result in local depletion of diversity. However, if this were the case, we would expect to see similar reductions in diversity across the autosomes, which we do not observe. Nonetheless, removal of closely related individuals in the

D. melanogaster Raleigh population (**Methods**) still results in depleted nucleotide diversity on the X relative to autosomes.

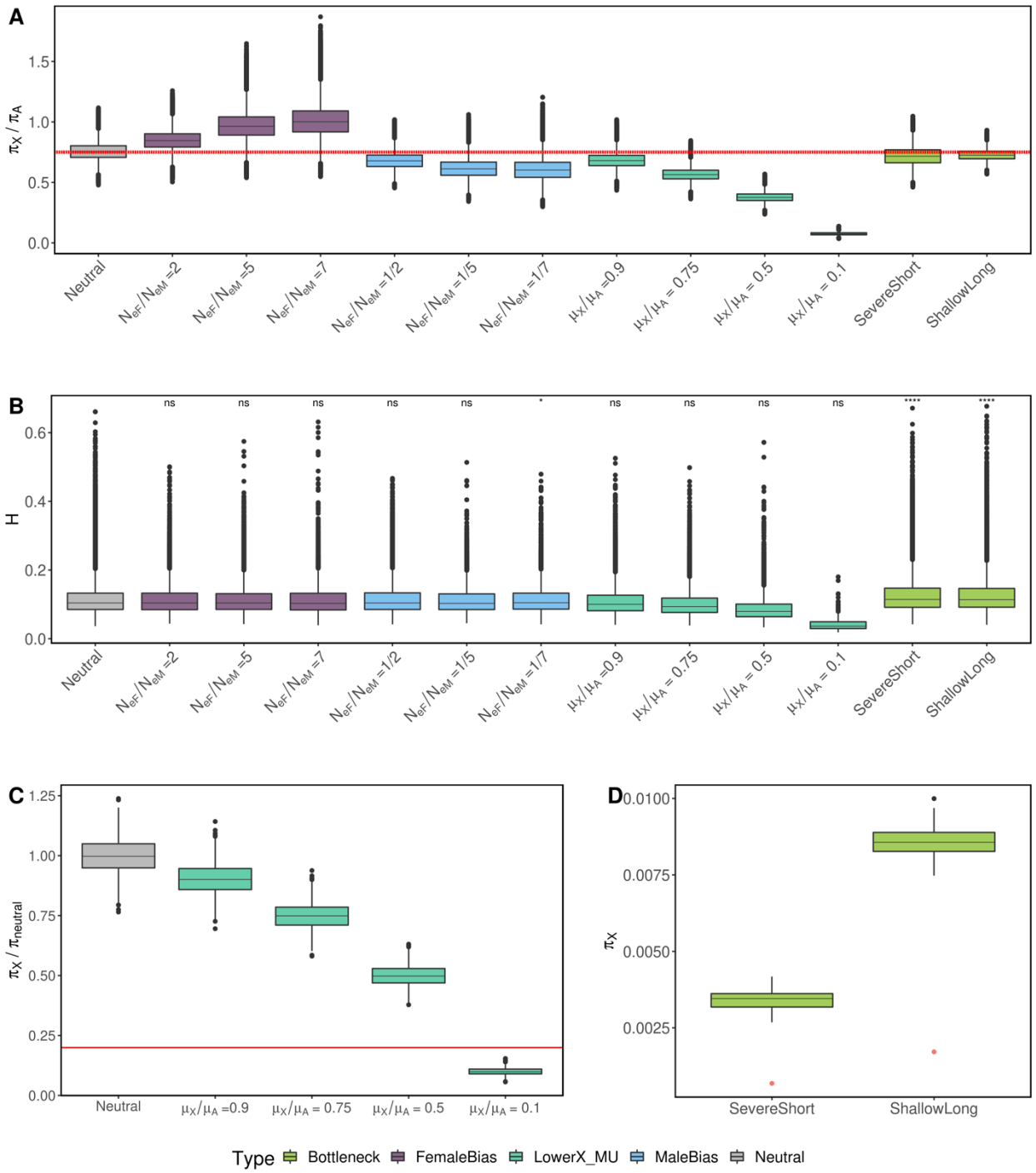

Figure 4. Nucleotide diversity (π) and haplotype homozygosity across 12 models with no selection. The models considered include three female-biased models (N_{eF}/N_{eM}=2, N_{eF}/N_{eM}=5, and

N_{eF}/N_{eM}=7), three male-biased models (N_{eM}/N_{eF}=2, N_{eM}/N_{eF} =5, and N_{eM}/N_{eF} = 7), four lower X-linked mutation rate models (μ_X/μ_A =0.9, μ_X/μ_A=0.75, μ_X/μ_A=0.5, and μ_X/μ_A=0.1, where μ_X is the X-linked mutation rate and μ_A is the autosomal mutation rate), and two bottleneck models (a severe short bottleneck and a shallow and long bottleneck). **(A)** π_X/π_A across the 12 models, where the red dashed line corresponds to the expected πX/πA = 0.75 value in the baseline neutral model (Same sex ratio, equal X and autosomal mutation rates and a constant N_e=10^6). **(B)** Haplotype homozygosity across all models. The asterisks represent a significant elevation in haplotype homozygosity relative to complete neutrality using a one-sided Wilcoxon rank-sum test. **(C)** $\pi_X/\pi_{neutral}$ across low X mutation rate models, with $\pi_{neutral}$ equal to the average π in the baseline neutral model. The horizontal red line is the low diversity threshold, defined as 20% of the average π/bp of the neutral case. **(D)** π_X in two bottleneck models. The red data point below the boxplots corresponds to 20% of the average π from each bottleneck model. We performed a total of 1,000 SLiM simulations for each scenario.

Evidence that hard sweeps can generate the patterns observed on the X across species

Having investigated neutral evolutionary scenarios, we next examined whether selection scenarios can generate patterns observed in the data. Both background selection (BGS) and selective sweeps can increase the proportion of singletons and decrease genetic diversity at linked sites, either through the purging of neutral alleles that are linked to deleterious mutations or through hitchhiking, in which a beneficial mutation and its genetic background spread rapidly through the population (44, 45, 65, 66).

To test whether BGS can, on its own, give rise to values of π_X/π_A<0.75, decrease π/bp below the low diversity threshold and elevate haplotype homozygosity, we simulated a population (Ne=1x10^6) in which a fraction (d=0.1,0.5 and 0.8) of mutations were deleterious. The selection coefficient (s_d) for the deleterious mutations followed a gamma distributed DFE with mean and shape parameter -0.000133 and 0.35, respectively (67). Our simulations show that (i) BGS does not decrease π_X such that π_X/π_A is considerably below 0.75 (**Fig. 5A**) (ii) only when 80% of mutations are deleterious diversity is reduced below 20% of $\pi_{neutral}$ (**Fig. 5B**) and (iii) none of the

BGS models considered can elevate homozygosity (**Fig. 5C**), consistent with previous findings (68, 69).

To test the combined effect of BGS and lower mutation rates on the X, we investigated the expected reduction in π in a region subject to strong versus weak BGS as a function of a constant deleterious selection strength s_d, deleterious mutation rate μ_d, and recombination rate ρ. Once again, we found that a window subject to strong BGS requires both 80% of mutations to be deleterious and a 10-fold increase in μ_d to generate dips in diversity below 20% of the average π/bp, similar to results when investigating the effects of BGS in isolation (**Text S1** and **Fig. S4**).

Next, we tested the effect of positive selection in generating dips in diversity and elevated haplotype homozygosity on the X chromosome. To do so, we simulated hard and soft sweeps (**Methods**), varied the strength of selection ($N_e s_b$ = 20, 200 and 2000) and computed the reduction of π/bp relative to neutrality ($\pi_{neutral}$) in a 20Kb window for each scenario. Our simulations show that selective sweeps can decrease π_X/π_A below 0.75 and elevate haplotype homozygosity (**Fig. 5A, C**), but only hard sweeps can reduce diversity below 20% of $\pi_{neutral}$ (**Fig. 5B**) as long as selection is sufficiently strong (Nes_b =2000). Thus, we conclude that patterns of reduced diversity and elevated haplotype homozygosity on the X chromosome of the six species analyzed are most consistent with hard sweeps.

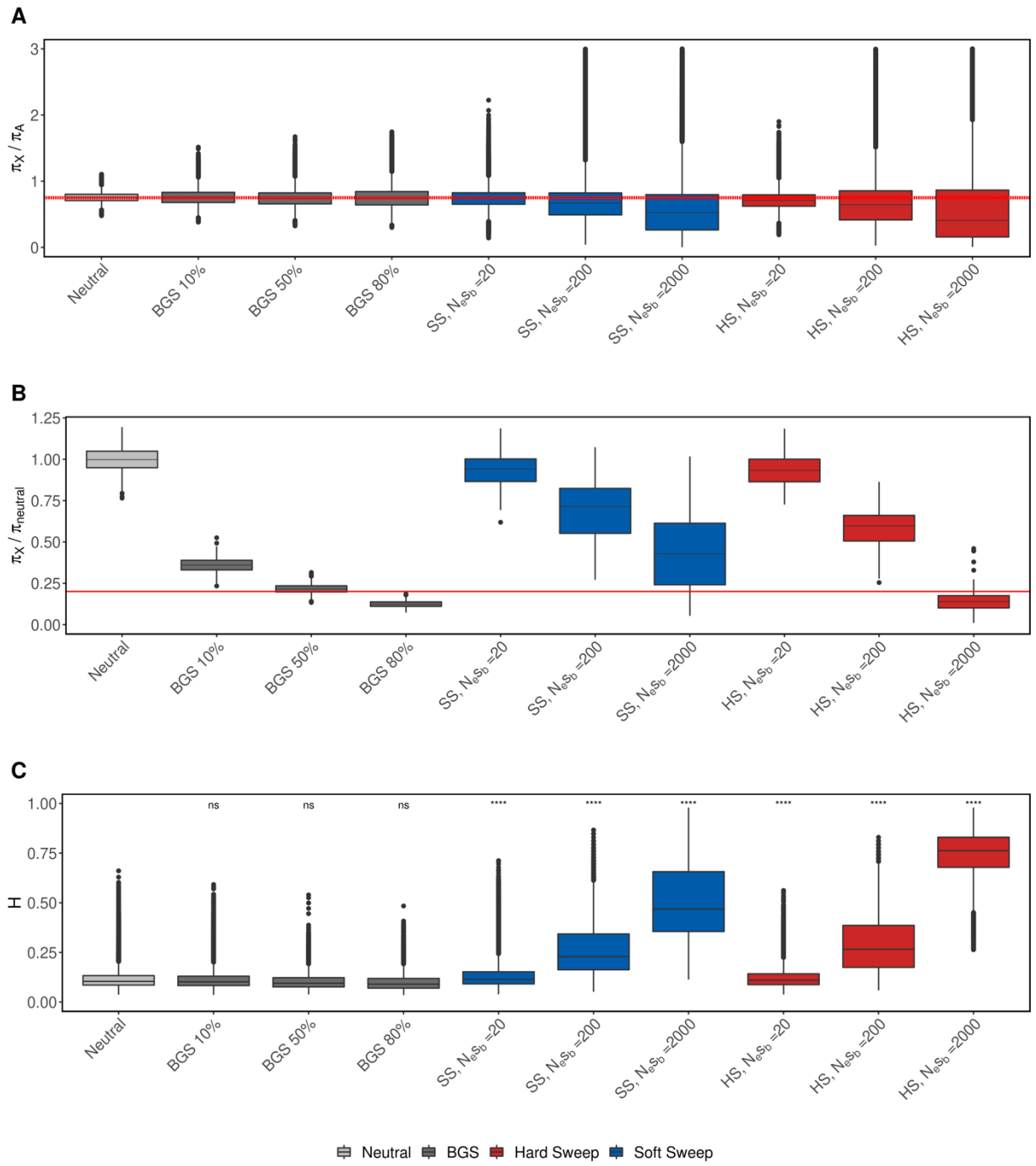

Figure 5. Background selection and hard selective sweeps reduce π below 20% of the neutral diversity but only selective sweep simulations can increase haplotype homozygosity. The models considered include three BGS (dark gray), three soft sweep (SS) and three hard sweep models (HS). For the BGS models, we varied the proportion of deleterious mutations (10, 50 and 80%) while the DFE for the deleterious selection coefficient (s_d) was gamma distributed with mean and shape parameter -0.000133 and 0.35, respectively (67). We simulated soft sweep (blue) and

hard sweep (red) models (Methods) varying the selection strength of the adaptive mutation ($N_e s_b$=20, 200, 2000). For each model, we computed: **(A)** π_X/π_A, where the red dashed line corresponds to the expected $\pi X/\pi A = 0.75$ value in a completely neutral case; **(B)** $\pi_X/\pi_{neutral}$ where $\pi_{neutral}$ is the average π in the baseline neutral model and the solid red line is 20% of $\pi_{neutral}$; and **(C)** haplotype homozygosity. The asterisks represent a significant elevation in haplotype homozygosity relative to complete neutrality using a one-sided Wilcoxon rank-sum test. We performed a total of 1,000 SLiM simulations for each scenario.

Discussion

A classic question in evolutionary biology is how the evolution of the X chromosome differs from that of autosomes given the X chromosome's central role in speciation, brain function, fertility, and sexual dimorphism (1–4, 9). Past work has suggested that the X chromosome may exhibit different evolutionary dynamics from the autosomes due to its unique inheritance pattern and increased exposure to selection through male hemizygosity (11, 12, 21, 22, 70). Recently, we found evidence in a North American *D. melanogaster* population that the X chromosome experiences an enrichment of hard selective sweeps compared to autosomes due to the increased visibility of new deleterious mutations to natural selection on the hemizygous X of males and reduced effective population size on the X (21). However, it is unclear whether this pattern of enrichment of hard sweeps on the X is a universal feature across all heterogametic species.

To understand if the enrichment of hard sweeps on the X is common across species, we analyzed multiple whole genome sequences from six *Drosophila* species and found that hard sweeps are in fact more common on the X chromosome than autosomes across these species. Additionally, we found that the observed patterns of diversity are inconsistent with soft sweeps, background selection, differences in mutation rate, population bottlenecks, or sex bias. This finding aligns with recent theoretical work predicting harder sweeps on the X chromosome (21, 70), as well as empirical work in apes (22) and *D. mauritiana* (64). In our analysis, we employed a wide range of statistics, including single-site (e.g. π/bp) and multi-site (e.g. haplotype homozygosity) statistics that are sensitive to signatures of selection and also leverage the additional resolution that whole genome sequences provide over genotype data. Combining single site and multi-locus statistics is still a relatively novel area of work (24, 27, 71–74) and has the potential to reveal patterns of evolution that cannot be detected by either type of statistic alone. Our approach

allows us to disentangle the effects of positive selection from other forces, such as background selection and demographic processes, and rule out inconsistent models.

Comparative population genetics, in which multiple genomes from several species each are considered simultaneously, is also a relatively new area of work given the paucity of deep population genetic samples from multiple species. Population genetic studies have traditionally focused on analyzing multiple genomes from a single species, e.g. (30, 75, 76) with few examining more than two population genetic whole genome datasets from different species simultaneously ((22, 77–79) with (79) restricted to exomes), leaving open numerous avenues of inquiry on the commonalities and idiosyncrasies of population genetic processes.

Now, with the increasing availability of deep population genetic sequences from multiple species, we can examine population genetic processes across many species. In this study, the ability to compare multiple species reveals that the extent of positive selection on the X chromosome may not be equal across all species. For example, we did not observe a higher proportion of low diversity windows on the X in *D. teissieri*, while in *D. santomea* we observed a low diversity region extending up to ~200Kb (**Fig. 2**). Additionally, this comparative approach reveals deviations from the trend in individual species. One species that in particular looked very different was *D. sechellia*. This species showed a unique trend in which π_X/π_A was the highest in exons and lowest in intergenic regions, even after stringent filtering of the data (**Methods**; **Figure 1**). A potential explanation for this behavior could be abundant introgression from *D. simulans* to *D. sechellia*, as previously reported (37–39). Interestingly, introgression appears to be less common on the X chromosome, potentially due to the involvement of sex chromosomes in hybrid incompatibilities (39, 42, 80). Moreover, functional regions of the genome are less likely to exhibit evidence of introgression, with exons being the least susceptible and intergenic regions having higher rates of introgressed regions (81). As a result, if introgression is abundant, the difference in diversity between the X chromosome and autosomes should be more pronounced in intergenic regions and least in exons, potentially driving the observed trend in the data.

Despite individual species showing deviations from the trend, most of the species analyzed in this study exhibited a higher proportion of low diversity regions on the X chromosome compared to the autosomes, suggesting more commonalities than differences across species. Towards this end, we found some targets of selection to be shared across species, with a higher proportion of low diversity windows than high diversity windows with orthologous gene groups overlapping

across two or more species (**Fig S5-S7**, **Text S2**). Nonetheless, despite several instances of shared ortholog groups across species, most genes found in low diversity windows were found in a single species only, indicating that while hard sweeps might be common on the X, different functions may be under selection in different species.

We acknowledge that the models, especially demographic models, assessed here are relatively simple compared to reality. However, they provide the simplest explanation for patterns of diversity observed in the data. It is possible that more complex models that incorporate a combination of evolutionary forces, excluding hard sweeps, can generate the observed dips in diversity and elevated haplotype homozygosity observed on the X. However, it is unclear what model this may be, and it may be difficult to identify such a model, especially a model that does not include any selection and simultaneously fits all summary statistics in the data (24).

Our paper is a clear demonstration of how comparative population genetic analyses can reveal new insights into the forces that shape genetic variation within and across species, which we expect to serve as a useful example as new high-resolution population genomic datasets from multiple species become increasingly available in the coming years. While our study did not examine the prevalence of soft sweeps in autosomes and the X chromosome due to limited samples sizes per species, future research with larger sample sizes could provide the relevant data needed to be able to detect soft sweeps with haplotype homozygosity statistics, thereby providing a more comprehensive understanding of the tempo and mode of adaptation across species (82). Ultimately, our study highlights the significance of hard sweeps in shaping the diversity patterns of the X chromosome across species and suggests an important evolutionary mechanism that may be widespread among all species. Future work may reveal the potential role of these hard sweeps in driving sexual dimorphism and speciation given the X chromosome's significant involvement in these important processes.

Methods

Data processing

Data for both *D. melanogaster* populations are publicly available as part of the Drosophila Genome Nexus data set (83) and were downloaded from www.johnpool.net. We downloaded 205 DGRP genomes from Raleigh (RAL), North Carolina and 197 DPGP3 genomes from Zambia (ZI).

For our analysis we used 100 genomes from each population. This data was previously processed (21) to remove individuals having high IBD with another , as well as residual heterozygosity, and high levels of missing data.

We downloaded *D. simulans* data from https://zenodo.org/record/154261#.YzMzty2z3jC consisting of 170 inbred lines from a North American population (76). For the remaining four species, we obtained genomes from NCBI's RefSeq and short read data from NCBI's Short Read Archive (42, 64, 84, 85) (**Table S3**).

Variant calling was performed with the NVIDIA Clara Parabricks pipeline v4.0.0 (https://docs.nvidia.com/clara/parabricks/4.0.0/index.html), a re-implementation of variant calling tools including BWA-mem (86) and GATK4 (87) optimized for running on NVIDIA graphics processing units. Following the GATK best practices for germline variant calling (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651), we mapped each sample's reads to the appropriate reference genome, sorted and removed PCR duplicates, generated single-sample GVCFs, then performed joint genotyping with HaplotypeCaller. From this initial set of variant calls, we removed sites with quality scores QUAL<30.0 to obtain a bootstrap set of high-confidence calls for another round of variant calling (https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-). We performed base quality score recalibration on the mapped reads using the bootstrap set then once again generated single sample GVCFs that were used for joint genotyping with HaplotypeCaller to ultimately generate a VCF file for each species. To remove any erroneous genotypes, we applied GATK recommended hard filters (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants) to our data and excluded sites that were not uniquely mappable. Additionally, we masked sites lying within repetitive elements as predicted by RepeatMasker (88). Only bi-allelic SNPs were considered in our analysis (**Fig S8**).

To mitigate potential confounding effects of selection, we calculated the proportion of base pairs identified as repeats within 50Kb windows. We removed windows in which 20% or more of the sites were marked as repeats by RepeatMasker. These high-density repeat regions are often associated with centromeric regions, which consist of highly homogeneous tandem repeats and are known to experience low rates of recombination (89–91). Consequently, these regions tend to exhibit reduced diversity and increased homozygosity, which can lead to false positive signals.

Diversity statistics and haplotype homozygosity

For all species except *D. simulans* we annotated SNPs using RefSeq (92) annotations downloaded at https://ftp.ncbi.nlm.nih.gov/genomes/refseq/. The corresponding assembly accession numbers for each species are: GCF_004382145.1 (*D. mauritiana)*, GCF_004382195.2 (*D. sechellia*), GCF_016746235.2 (*D. teissieri*), GCF_016746245.2 (*D. santomea*) and GCF_000001215.2 (*D. melanogaster*). For *D. simulans* we used the reference genome and corresponding gff file provided by Rebekah Rogers and Peter Andolfatto.

Next, we divided the data into exon introns and intergenic regions using BEDtools v.2.3.0 (93). If any position could be included in multiple categories, we annotated the position prioritizing exons followed by introns and finally intergenic regions. Finally, we excluded features with less than 40bp.

Then, we calculated π_X and π_X in each of the exons, introns and intergenic regions obtained previously. To compute π_X/π_A we performed a total of 1,000 bootstrap replicates in which we sampled exons introns and intergenic regions at random, computing the mean π_X, mean π_A and mean(π_X)/mean(π_A) in each sample. We approximated π by $\pi \approx 2pq$, with p the frequency of the major allele at a given site and q=1-p.

We also computed π/bp in non-overlapping windows of 10, 20 and 50 kb (**Fig.1, Fig S3**) across the autosomes and the X chromosome of all species. We then defined a low diversity threshold as X% of the π/bp average of each chromosome with X=20 or 30%. Windows with π/bp below this threshold were labelled as low diversity windows. For the main analysis of this work, 20Kb windows and a threshold of 20% of the average π/bp was used.

In addition to nucleotide diversity, we computed haplotype or multi-locus genotype homozygosity in windows of length 20 SNPs long on the X chromosome. For species for which we had phased data (*D. melanogaster* and *D. simulans*) we computed haplotype homozygosity. For the remaining species we computed multi-locus genotype homozygosity. We also tested windows of length 50 and 10 SNPs long. However, due to the small sample size of most species, the probability of observing the same multi-locus genotype twice becomes small with longer window sizes, making it difficult to capture any signal in the 50 SNP window case. With 10 SNP windows we also found an elevation of homozygosity in low diversity windows. However, we opted for 20 SNP windows as smaller window sizes may increase homozygosity due to random

drift. Importantly, we kept the window size constant across low versus high diversity windows in order to make these two regions comparable.

Simulation analysis

We simulated a variety of evolutionary models for the X chromosome and autosomes using SLiM 3.7. Excluding the bottleneck models, we simulate a constant $Ne=10^6$ population. An N_e of this order of magnitude is reasonable for most of the species analyzed given mean nucleotide diversity levels, with exception of *D. sechellia*, which has been shown to have a lower $Ne \sim 10^5$ (94). However, SLiM is a forward in time simulator, which makes simulating populations with $N_e > 5x10^5$ unfeasible due to memory requirements. As in our previous work, we rescaled our simulations using a constant factor of Q=50.

First, we simulated scenarios that evolve neutrally. Next, we simulate selective scenarios including background selection, hard sweeps and soft sweeps, as described below. We compute π/bp and haplotype homozygosity in all scenarios to evaluate whether any of the models could generate the patterns observed in the data.

In all our simulations we modeled a 20kb region with a recombination rate of $\rho=5x10^{-7}$ cM/bp and a neutral mutation rate of $\mu=1x10^{-9}$ rescaled by Q=50. For the simulations that include selection we assumed that mutations on the X of males experience the same fitness effect that a homozygous female (i.e. dosage compensation). We ran a total of 1,000 simulations for each model and ran a $10N_e$ burn-in for every simulation.

Neutral models

We first simulated a completely neutral model with a female to male ratio of one and equal mutation rates between the sexes. Next, we simulated female and male sex bias varying the sex ratio across 2:1, 5:1 and 7:1 for each scenario. We also simulated a model with a lower X-linked mutation rate reducing the X-linked mutation rate such that the ratio μ_X/μ_A was 0.1,05, 0.75 or 0.9.

The expected lower population size of the X can result in stronger drift that can be exacerbated in a bottleneck. For this reason, we considered two bottleneck models that were fit to to π/bp and S/bp in short introns from DGRP dataset (24, 32) : (i) a severe and short bottleneck with a bottleneck N_e of $0.002N_{e,ancestral}$ for $0.0002*2N_{e,ancestral}$ generations, and (ii) a shallow and long bottleneck with a bottleneck N_e of $0.4N_{e,ancestral}$ for $0.056*2N_{e,ancestral}$ generations.

Background selection

Next, we simulated a constant $Ne=10^6$ with deleterious variation. The selection coefficients for deleterious mutations were gamma distributed with mean and shape parameter -0.000133 and 0.35, respectively (Huber et al. 2017). We varied the percentage of deleterious mutations simulating the case in which 10, 50 or 80% of incoming mutations were deleterious.

Selective sweeps

We simulated soft sweeps by introducing adaptive mutations recurrently at the center of the haplotype at a rate defined by $\theta_a=1$, where $\theta_a=4Ne\mu_a$ with μ_a the adaptive mutation rate. We conditioned our simulations on the fixation of the sweep, after which we took a sample of n=100 haplotypes. Soft sweeps are expected for $\theta_a>0.1$ (17, 19), however we verified that our sample represented a soft sweep by only including samples with two or more mutational origins in our analysis.

To model hard sweeps, we introduced a single adaptive mutation on the center of the haplotype and restarted the simulation if the adaptive mutation was lost. Again, we conditioned on fixation of the sweep and took a sample of n=100 haplotypes. For both hard and soft sweeps, we varied the strength of selection such that Nes_b=20, 200 or 2000.

Orthologs in windows below and above the low diversity threshold

To identify shared regions under selection, we obtained the genes intersecting windows below and above the low diversity threshold using BEDtools v.2.3.0 and the annotation files from the reference genome for each species in our study. Next, we used the gene IDs and matched them to their corresponding ortholog IDs in orthoDB v11 (95). We then looked for overlapping orthologs across species and obtained the proportion of orthologs that overlap over two, three or four species in low versus high diversity windows.

To obtain windows above the threshold that were comparable to those below the threshold we obtained 100 random samples of *n* orthologs from high diversity regions with *n*= the number of orthologs in low diversity regions per species. From this we computed the mean proportion of orthologs as well as the 95% confidence interval in high diversity regions.

Code availability

Code used for simulations as well as to process and analyze the data is available at https://github.com/garudlab/DrosCrossSpecies_XchrHardSweeps

Acknowledgments

We are grateful to Rebekah Rogers and Peter Andolfatto for providing the *D.simulans* w501 genome assembly. We thank Dmitri Petrov and Kirk Lohmueller for helpful conversations. MH was supported by the National Institutes of Health Systems and Integrative Biology Training Grant (NIH-NIGMS 5T32GM008185) and the National Institutes of Health Training Grant in Genomic Analysis and Interpretation (NIH T32HG002536). BYK was supported by NIH grant NIGMS F32GM135998. NRG was supported by funding from a National Science Foundation CAREER award.

References

1. B. A. Payseur, D. C. Presgraves, D. A. Filatov, Introduction: Sex chromosomes and speciation in *Molecular Ecology*, (2018), pp. 3745–3748.
2. G. M. Saifi, H. S. Chandra, An apparent excess of sex and reproduction-related genes on the human X chromosome. *Proceedings of the Royal Society B: Biological Sciences* **266**, 203–209 (1999).
3. W. R. Rice, Sex chromosomes and the evolution of sexual dimorphism. *Evolution (N Y)* **38**, 735–742 (1984).
4. D. H. Skuse, X-linked genes and mental functioning. *Hum Mol Genet* **14**, 27–32 (2005).
5. M. Turelli, H. A. Orr, The dominance theory of HALDANE's rule. *Genetics* **140**, 389–402 (1995).
6. J. A. Coyne, The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **353**, 287–305 (1998).
7. M. M. Patten, The X chromosome favors males under sexually antagonistic selection. *Evolution (N Y)* **73**, 84–91 (2019).
8. A. Glaser-Schmitt, M. J. Wittmann, T. J. S. Ramnarine, J. Parsch, Sexual Antagonism, Temporally Fluctuating Selection, and Variable Dominance Affect a Regulatory Polymorphism in Drosophila melanogaster . *Mol Biol Evol* **38**, 4891–4907 (2021).
9. R. Dean, J. E. Mank, The role of sex chromosomes in sexual dimorphism: Discordance between molecular and phenotypic data. *J Evol Biol* **27**, 1443–1453 (2014).
10. B. Charlesworth, J. A. Coyne, N. H. Barton, The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist* **130**, 113–146 (1987).

11. B. Vicoso, B. Charlesworth, Evolution on the X chromosome: Unusual patterns and processes. *Nat Rev Genet* **7**, 645–653 (2006).
12. B. Charlesworth, J. L. Campos, B. C. Jackson, Faster-X evolution: Theory and evidence from Drosophila. *Mol Ecol* **27**, 3753–3771 (2018).
13. B. Vicoso, B. Charlesworth, Effective population size and the faster-X effect: An extended model. *Evolution (N Y)* **63**, 2413–2426 (2009).
14. H. A. Orr, A. J. Betancourt, Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157**, 875–884 (2001).
15. J. Maynard Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet Res (Camb)* **23**, 23–35 (1974).
16. N. L. Kaplan, R. R. Hudson, C. H. Langley, The "hitchhiking effect" revisited. *Genetics* **123**, 887–899 (1989).
17. J. Hermisson, P. S. Pennings, Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).
18. J. Hermisson, P. S. Pennings, Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol Evol* **8**, 700–716 (2017).
19. P. S. Pennings, J. Hermisson, Soft sweeps II - Molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* **23**, 1076–1084 (2006).
20. P. W. Messer, D. A. Petrov, Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28**, 659–669 (2013).
21. M. Harris, N. R. Garud, Enrichment of Hard Sweeps on the X Chromosome in Drosophila melanogaster. *Mol Biol Evol* **40**, 1–16 (2023).
22. K. Nam, *et al.*, Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc Natl Acad Sci U S A* **112**, 6413–6418 (2015).
23. R. B. Harris, A. Sackman, J. D. Jensen, On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PLoS Genet* **14**, e1007859 (2018).
24. N. R. Garud, P. W. Messer, D. A. Petrov, Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data. *PLoS Genet* **17** (2021).
25. A. F. Feder, P. S. Pennings, D. A. Petrov, The clarifying role of time series data in the population genetics of HIV. *PLoS Genet* **17** (2021).
26. B. M. Peter, E. Huerta-Sanchez, R. Nielsen, Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS Genet* **8**, e1003011 (2012).
27. D. R. Schrider, A. D. Kern, S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genet* **12** (2016).
28. D. R. Schrider, F. K. Mendes, M. W. Hahn, A. D. Kern, Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**, 267–284 (2015).
29. Z. J. Assaf, D. A. Petrov, J. R. Blundell, Obstruction of adaptation in diploids by recessive, strongly deleterious alleles. *Proc Natl Acad Sci U S A* **112**, E2658–E2666 (2015).
30. L. Arbiza, S. Gottipati, A. Siepel, A. Keinan, Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet* **94**, 827–844 (2014).
31. C. McGrath, Highlight: Comparative Population Genomics—Answering Old Questions with New Data. *Genome Biol Evol* **14** (2022).

32. N. R. Garud, P. W. Messer, E. O. Buzbas, D. A. Petrov, Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet* **11**, 1–32 (2015).
33. N. D. Singh, J. M. Macpherson, J. D. Jensen, D. A. Petrov, Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of Drosophila melanogaster. *BMC Evol Biol* **7**, 1–16 (2007).
34. J. E. Pool, *et al.*, Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLoS Genet* **8** (2012).
35. M. Kauer, B. Zangerl, D. Dieringer, C. Schlötterer, Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster. *Genetics* **160**, 247–256 (2002).
36. G. McVicker, D. Gordon, C. Davis, P. Green, Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5** (2009).
37. D. R. Matute, J. F. Ayroles, Hybridization occurs between Drosophila simulans and D. sechellia in the Seychelles archipelago. *J Evol Biol* **27**, 1057–1068 (2014).
38. D. Garrigan, *et al.*, Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res* **22**, 1499–1511 (2012).
39. D. R. Schrider, J. Ayroles, D. R. Matute, A. D. Kern, Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. Sechellia. *PLoS Genet* **14** (2018).
40. C. A. Muirhead, D. C. Presgraves, Hybrid incompatibilities, local adaptation, and the genomic distribution of natural introgression between species. *American Naturalist* **187**, 249–261 (2016).
41. C. Fraïsse, H. Sachdeva, The rates of introgression and barriers to genetic exchange between hybridizing species: Sex chromosomes vs autosomes. *Genetics* **217** (2021).
42. D. A. Turissini, D. R. Matute, Fine scale mapping of genomic introgressions within the Drosophila yakuba clade. *PLoS Genet* **13** (2017).
43. L. S. Gramates, *et al.*, FlyBase: a guided tour of highlighted features. *Genetics* **220** (2022).
44. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
45. D. Charlesworth, B. Charlesworth, M. T. Morgan, The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632 (1995).
46. J. C. Fay, C. I. Wu, Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).
47. M. Degiorgio, C. D. Huber, M. J. Hubisz, I. Hellmann, R. Nielsen, SweepFinder2: Increased sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895–1897 (2016).
48. J. D. Wall, J. K. Pritchard, Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587–597 (2003).
49. A. M. Harris, N. R. Garud, M. Degiorgio, Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity. *Genetics* **210**, 1429–1452 (2018).
50. A. J. Betancourt, Y. Kim, H. A. Orr, A pseudohitchhiking model of X vs. autosomal diversity. *Genetics* **168**, 2261–2269 (2004).
51. B. C. Haller, P. W. Messer, SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Mol Biol Evol* **36**, 632–637 (2019).

52. D. Dieringer, V. Nolte, C. Schlötterer, Population structure in African Drosophila melanogaster revealed by microsatellite analysis. *Mol Ecol* **14**, 563–573 (2005).
53. K. Thornton, P. Andolfatto, Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics* **172**, 1607–1619 (2006).
54. J. B. Drost, W. R. Lee, Biological basis of germline mutation: Comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environ Mol Mutagen* **25**, 48–64 (1995).
55. M. Kirkpatrick, D. W. Hall, "Male-biased mutation, sex linkage, and the rate of adaptive evolution" (2004).
56. J. L. Campos, K. Zeng, D. J. Parker, B. Charlesworth, P. R. Haddrill, Codon usage bias and effective population sizes on the X chromosome versus the autosomes in drosophila melanogaster. *Mol Biol Evol* **30**, 811–823 (2013).
57. N. D. Singh, A. M. Larracuente, A. G. Clark, Contrasting the efficacy of selection on the X and autosomes in drosophila. *Mol Biol Evol* **25**, 454–467 (2008).
58. D. R. Schrider, D. Houle, M. Lynch, M. W. Hahn, Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics* **194**, 937–954 (2013).
59. B. Vicoso, P. R. Haddrill, B. Charlesworth, A multispecies approach for comparing sequence evolution of X-linked and autosomal sites in Drosophila. *Genet Res (Camb)* **90**, 421–431 (2008).
60. P. D. Keightley, R. W. Ness, D. L. Halligan, P. R. Haddrill, Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics* **196**, 313–320 (2014).
61. N. D. Singh, J. C. Davis, D. A. Petrov, X-linked genes evolve higher codon bias in Drosophila and Caenorhabditis. *Genetics* **171**, 145–155 (2005).
62. N. D. Singh, J. C. Davis, D. A. Petrov, Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. *J Mol Evol* **61**, 315–324 (2005).
63. P. D. Keightley, *et al.*, Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res* **19**, 1195–1201 (2009).
64. D. Garrigan, S. B. Kingan, A. J. Geneva, J. P. Vedanayagam, D. C. Presgraves, Genome diversity and divergence in Drosophila mauritiana: Multiple signatures of faster X evolution. *Genome Biol Evol* **6**, 2444–2458 (2014).
65. B. Charlesworth, M. T. Morgan, D. Charlesworth, The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
66. W. Stephan, Genetic hitchhiking versus background selection: The controversy and its implications. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1245–1253 (2010).
67. C. D. Huber, B. Y. Kim, C. D. Marsden, K. E. Lohmueller, Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci U S A* **114**, 4465–4470 (2017).
68. D. Enard, P. W. Messer, D. A. Petrov, Genome-wide signals of positive selection in human evolution. *Genome Res* **24**, 885–895 (2014).

69. D. R. Schrider, Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics* **216**, 499–519 (2020).
70. P. Muralidhar, C. Veller, Dominance shifts increase the likelihood of soft selective sweeps. *Evolution (N Y)*, 1–19 (2022).
71. S. Sheehan, Y. S. Song, Deep Learning for Population Genetic Inference. *PLoS Comput Biol* **12**, e1004845 (2016).
72. P. Johri, B. Charlesworth, J. D. Jensen, Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics* **215**, 173–192 (2020).
73. A. P. Ragsdale, R. N. Gutenkunst, Inferring demographic history using two-locus statistics. *Genetics* **206**, 1037–1048 (2017).
74. K. Lin, H. Li, C. Schlötterer, A. Futschik, Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics* **187**, 229–244 (2011).
75. N. R. Garud, N. A. Rosenberg, Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps. *Theor Popul Biol* **102**, 94–101 (2015).
76. S. A. Signor, F. N. New, S. Nuzhdin, A large panel of drosophila simulans reveals an abundance of common variants. *Genome Biol Evol* **10**, 189–206 (2018).
77. Z. H. Chen, *et al.*, Contrasting patterns of genomic diversity reveal accelerated genetic drift but reduced directional selection on X-chromosome in wild and domestic sheep species. *Genome Biol Evol* **10**, 1282–1297 (2018).
78. K. Nadachowska-Brzyska, R. Burri, H. Ellegren, Footprints of adaptive evolution revealed by whole Z chromosomes haplotypes in flycatchers. *Mol Ecol* **28**, 2290–2304 (2019).
79. T. Latrille, N. Rodrigue, N. Lartillot, Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proceedings of the National Academy of Sciences* **120** (2023).
80. L. S. Maroja, E. L. Larson, S. M. Bogdanowicz, R. G. Harrison, Genes with restricted introgression in a field cricket (Gryllus firmus/Gryllus pennsylvanicus) hybrid zone are concentrated on the X chromosome and a single autosome. *G3: Genes, Genomes, Genetics* **5**, 2219–2227 (2015).
81. S. Sankararaman, *et al.*, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
82. P. S. Pennings, J. Hermisson, Soft sweeps III: The signature of positive selection from recurrent mutation. *PLoS Genet* **2**, 1998–2012 (2006).
83. J. B. Lack, *et al.*, The drosophila genome nexus: A population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics* **199**, 1229–1241 (2015).
84. M. K. Meany, *et al.*, Loss of cytoplasmic incompatibility and minimal fecundity effects explain relatively low Wolbachia frequencies in Drosophila mauritiana. *Evolution (N Y)* **73** (2019).
85. A. Serrato-Capuchina, *et al.*, P-elements strengthen reproductive isolation within the Drosophila simulans species complex. *Evolution (N Y)* **75** (2021).
86. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* (2013).
87. G. Van der Auwera, B. O'Connor, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st Ed. (O' Riley Media, 2020).

88. A. F. A. Smit, R. Hubley, P. Grenn, RepeatMasker Open-4.0. *RepeatMasker Open* (2015).
89. N. Vincenten, *et al.*, The kinetochore prevents centromere-proximal crossover recombination during meiosis. *Elife* **4** (2015).
90. K. Mather, Crossing Over and Heterochromatin in the X Chromosome of Drosophila Melanogaster. *Genetics* **24**, 413–435 (1939).
91. R. P. Levine, Chromosome Structure and the Mechanism of Crossing Over. *Proceedings of the National Academy of Sciences* **41**, 727–730 (1955).
92. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
93. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
94. D. Legrand, *et al.*, Species-wide genetic variation and demographic history of Drosophila sechellia, a species lacking population structure. *Genetics* **182**, 1197–1206 (2009).
95. D. Kuznetsov, *et al.*, OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451 (2023).

Supplemental text to Enrichment of hard sweeps on the X chromosome compared to autosomes in six *Drosophila* species

Mariana Harris, Bernard Kim, and Nandita Garud

Table of Contents

Supplemental Tables:

Table S1	Drosophila species and populations used in study	Page 2
Table S2	Minimum and mean π/bp computed in non-overlapping 20Kb windows on the X chromosome across species.	Page 3

Supplemental Figures:

Figure S1	Genome-wide diversity in six species.	Page 4
Figure S2	Distribution of π/bp for the X chromosome across species.	Page 5
Figure S3	Low diversity windows on the X versus autosomes as a function of the low diversity thresholds and window size.	Page 6
Figure S4	Expected reduction in π from BGS.	Page 7
Figure S5	Proportion of orthologs shared across species in windows below and above the low diversity threshold	Page 8
Figure S6	Proportion of orthologs shared across species in windows below and above the low diversity threshold in *D.melanogaster* (ZI) and *D. melanogaster* (RA) only	Page 9
Figure S7	Number of orthologs shared across species.	Page 10
Figure S8	Phred-scaled quality score (QUAL) across species.	Page 11

Supplemental Texts:

Text S1	The effect of background selection on diversity	Page 12
Text S2	Overlapping orthologous gene groups with low diversity on the X chromosome across species	Page 13

Species	Clade	Population	# Samples used	Source
D. melanogaster	*melanogaster*	North America (RAL)	100	MacKay et al. 2012; Lack et al 2015
D. melanogaster	*melanogaster*	Africa (ZI)	100	Pool et al 2012; Lack et al 2015
D. simulans	*simulans*	North America (CA)	170	Signor 2018
D. sechellia	*simulans*	Africa (Praslin)	23	Schrider et al. 2018
D. mauritiana	*simulans*	Africa (Mauritius)	15	Garrigan et al. 2014 Serrato-Capuchina et al. 2021 Meany et al. 2019
D. santomea	*yakuba*	Africa (São Tomé)	7	Turissini and Matute 2017; Bachtrog et al 2007;
D. teissieri	*yakuba*	Africa (Bioko, Zimbabwe, Gabon)	11	Turissini and Matute 2017

Table S1. Drosophila species and populations used in study. Clade, population of origin, number of samples, and reference are given in each column.

Species	min(π/bp)	mean(π/bp)	% of mean(π/bp)
D.mauritiana	0.00004	0.006	0.7%
D. melanogaster (RA)	0.0002	0.003	6.7%
D. melanogaster (ZI)	0.0006	0.008	7.5%
D.santomea	0.0005	0.005	10.0%
D. sechellia	0.000002	0.0003	0.7%
D. simulans	0.00002	0.004	0.5%
D. teissieri	0.0003	0.01	3.0%

Table S2. Minimum and mean π/bp computed in non-overlapping 20Kb windows on the X chromosome across species. The fourth column shows the minimum π/bp on the X as a percent of the average π/bp on the X chromosome. In *D. simulans* the window with lowest diversity is 0.5% of the average π/bp while on *D. santomea* it is 10% of the average π/bp.

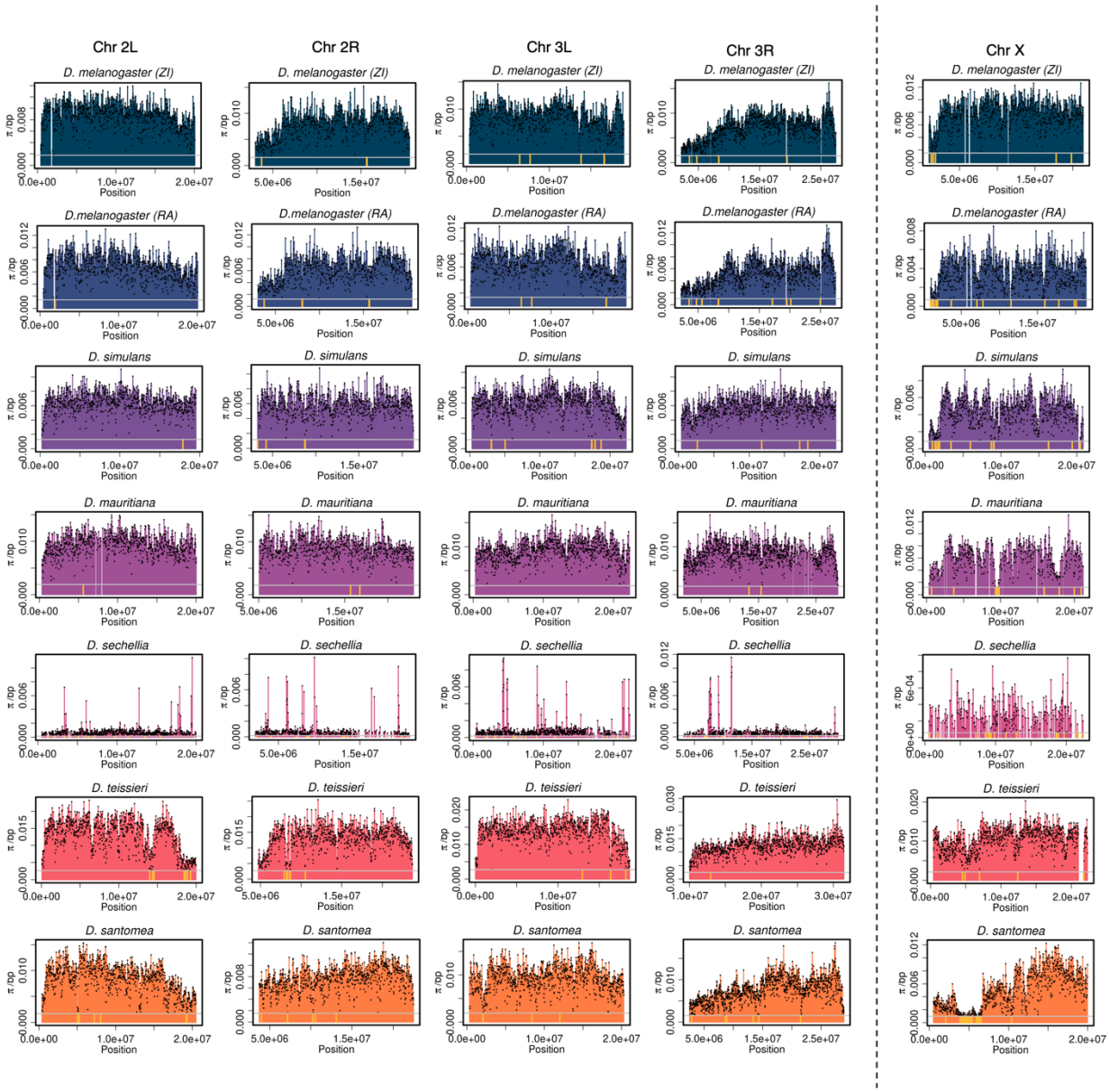

Figure S1. Genome-wide diversity in six species. π/bp in 20kb bins across four autosomal arms and the X chromosome of six *Drosophila* species and two *D. melanogaster* populations from Raleigh (RA) and Zambia (ZI). Windows where diversity is below 20% of the chromosomal average (grey horizontal line) are highlighted in gold.

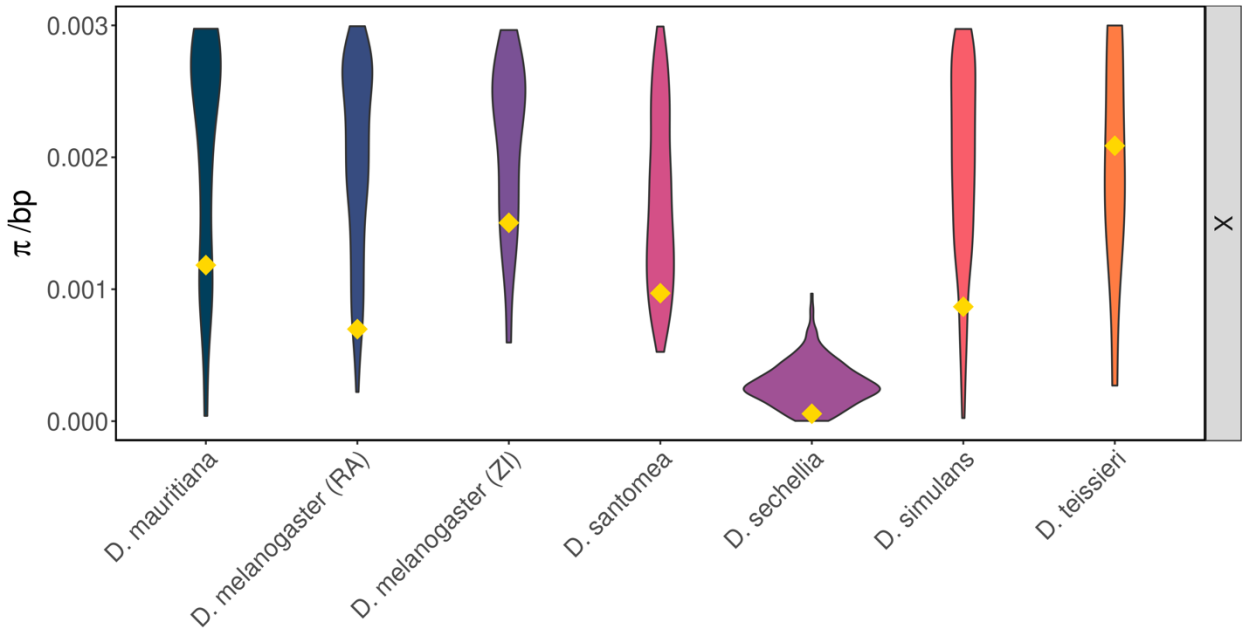

Figure S2. Distribution of π/bp for the X chromosome across species. Violin plots for π/bp for each species' X chromosome are plotted with y-axis going up to π/bp=0.03 for better visibility of the lower tail of the distribution. The low diversity threshold is shown as the yellow diamond on the tail of each violin plot and represents 20% of the mean value of the respective distribution. The lower tail of each distribution extends well below the low diversity threshold for all species.

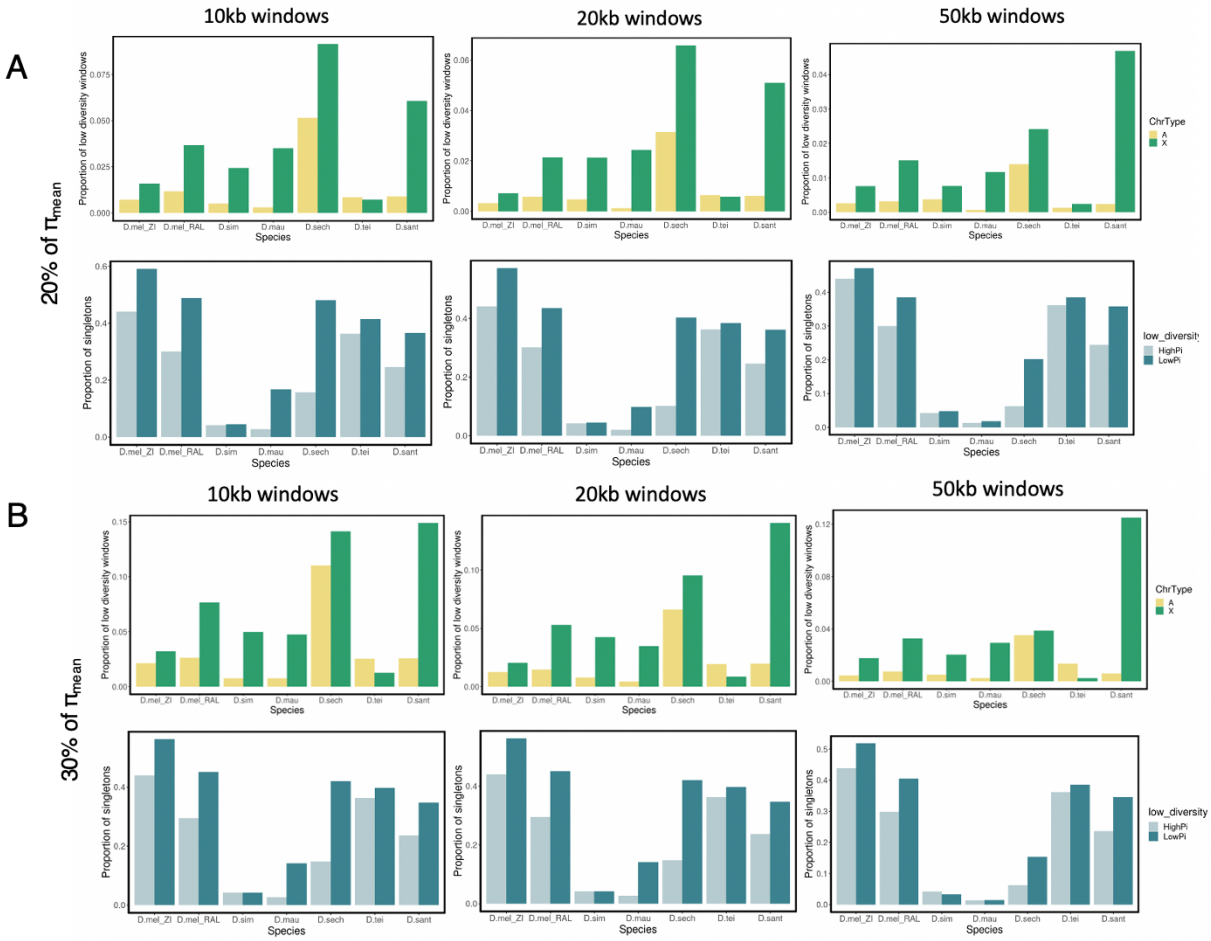

Figure S3. Low diversity windows on the X versus autosomes as a function of the low diversity thresholds and window size. Proportion of low diversity windows on X versus autosomes (top row, green and yellow bars) and proportion of singletons in low diversity windows on the X chromosome (bottom row, blue bars). Three different analysis windows were tested: 10kb, 20kb and 30kb windows, as well as two different diversity thresholds: **(A)** 20% and **(B)** 30% of the chromosome's π/bp average. We consistently see a higher proportion of low diversity windows on the X compared to autosomes. Additionally, these low diversity windows show a higher proportion of singletons that windows above the threshold. This figure is analogous to figures **Fig. 2B,C** in the main text.

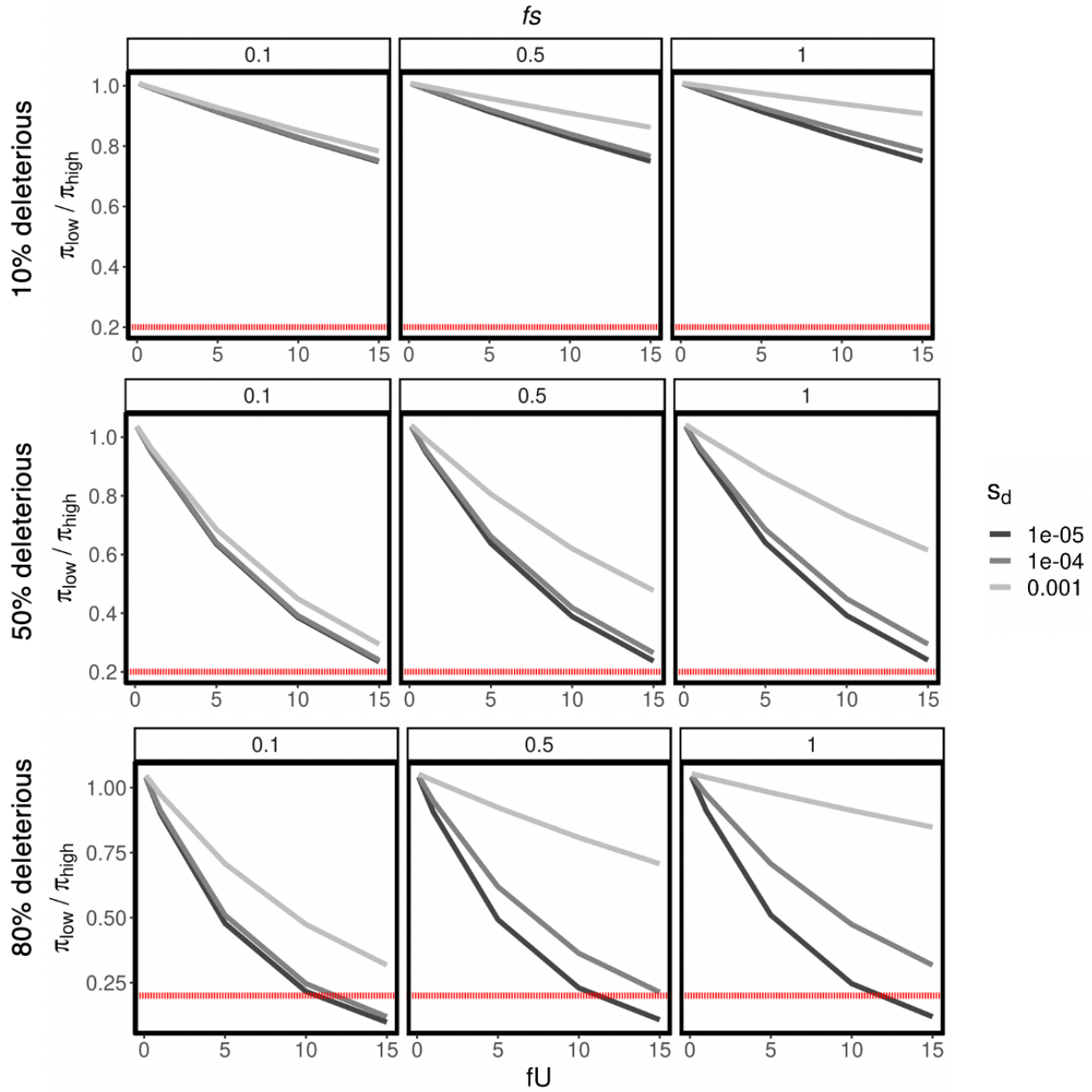

Figure S4. Expected reduction in π from BGS. Simulation strategy is described in detail in Text S1. Rows represent different proportions of deleterious mutations from the total incoming mutations per generation with 10,50 and 80% of deleterious mutations from top to bottom. The columns are the fraction of the outside deleterious selection strength s_d, so that selection at low diversity windows is $0.1s_d$,$0.5s_d$ and equal to s_d. The x-axis (f_U) is the relative mutation rate in low versus high diversity regions such that an an $f_U = 10$ means that the deleterious mutation rate is 10 times higher in the window subject to strong BGS. The different shades of gray lines correspond to different s_d values with the weakest being the darkest shade. Finally, the dashed red line represents 20% of the diversity of windows outside the high BGS regions.

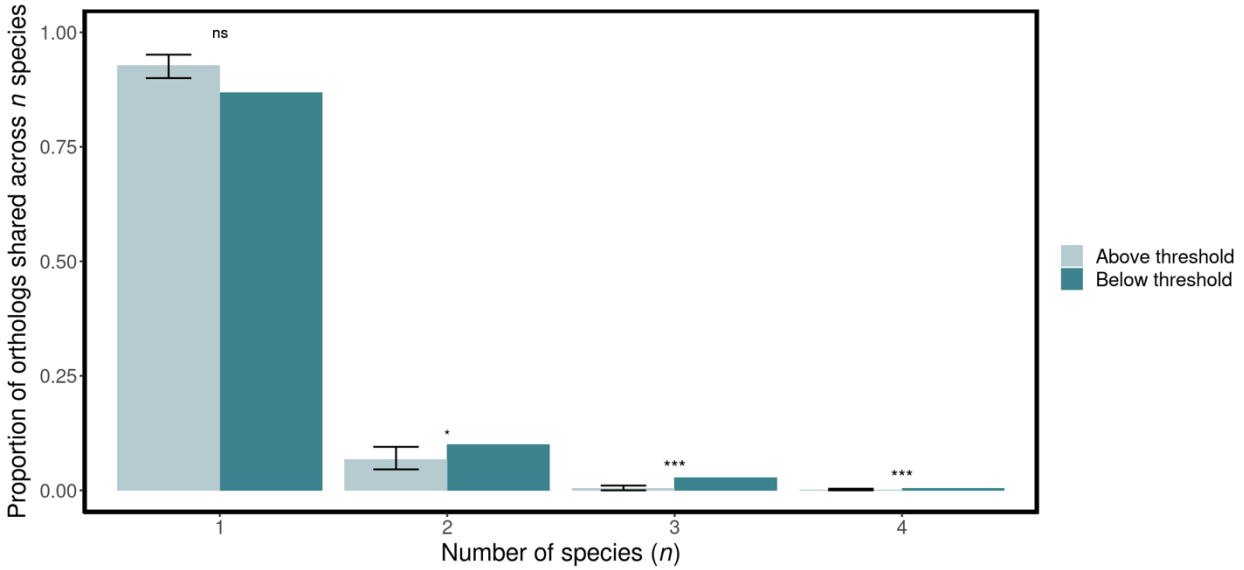

Figure S5. Proportion of orthologs shared across species in windows below and above the low diversity threshold. We obtained 100 random samples of orthologs found in windows above the low diversity threshold and obtained the mean number of orthologs shared across *n=1,2,3*or 4 species (light blue) as well as the corresponding 95% confident intervals for each case. The case of *n=1* represents orthologs found in a single species only. For *n>1,* we found that the proportion of orthologs shared across species is higher in low diversity windows. Comparisons between the two *D. melanogaster* populations were excluded from the analysis in order to emphasize between-species dynamics.

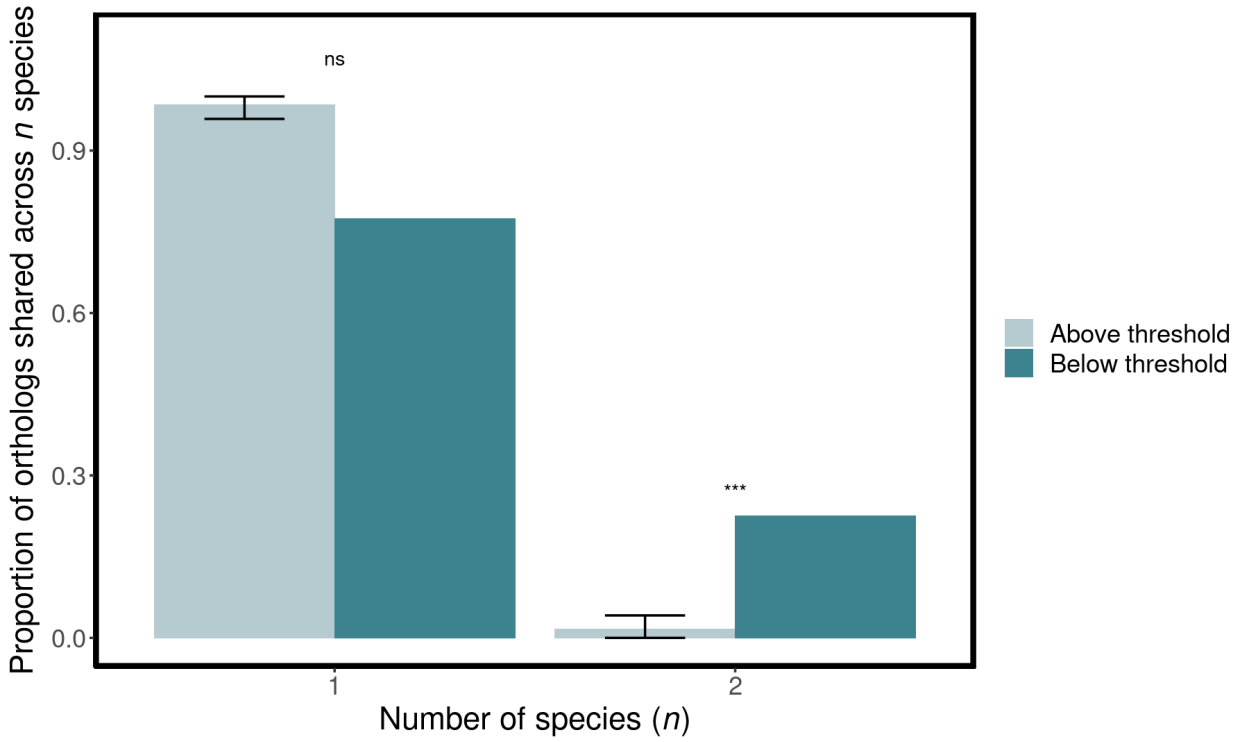

Figure S6. Proportion of orthologs shared across species in windows below and above the low diversity threshold in *D. melanogaster* (ZI) and *D. melanogaster* (RA) only. We obtained 100 random samples of orthologs found in windows above the low diversity threshold and obtained the mean number of orthologs shared across *D. melanogaster* populations (light blue) as well as the corresponding 95% confident intervals. We show the comparison between *D. melanogaster* populations in order to emphasize processes that may be influencing different populations of the same species.

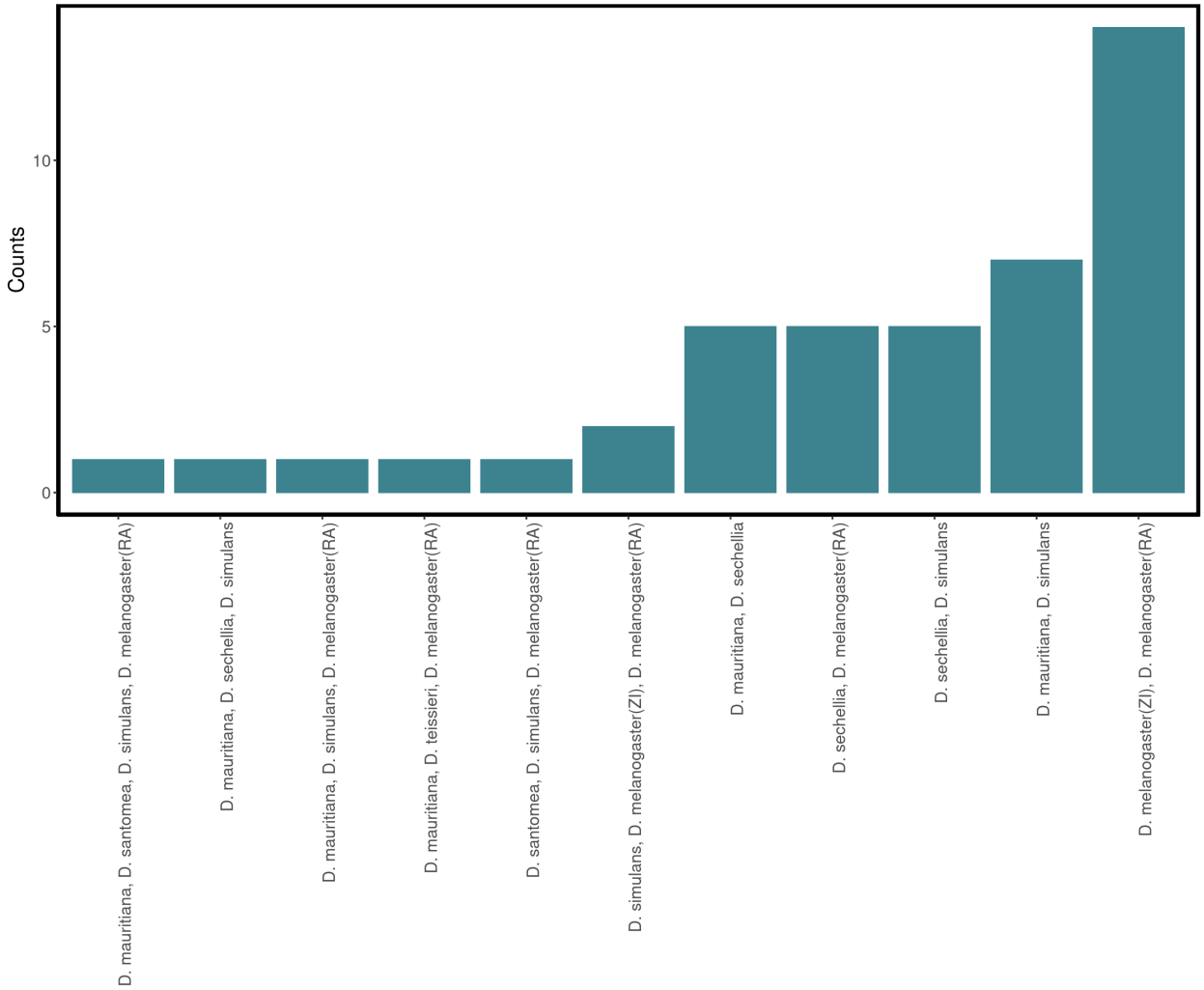

Fig S7. Number of orthologs shared across species. The x axis shows the clusters of species that have orthologs in common. The *D. melanogaster* populations have the highest number of overlapping orthologs. Overlap between three or more species is less common.

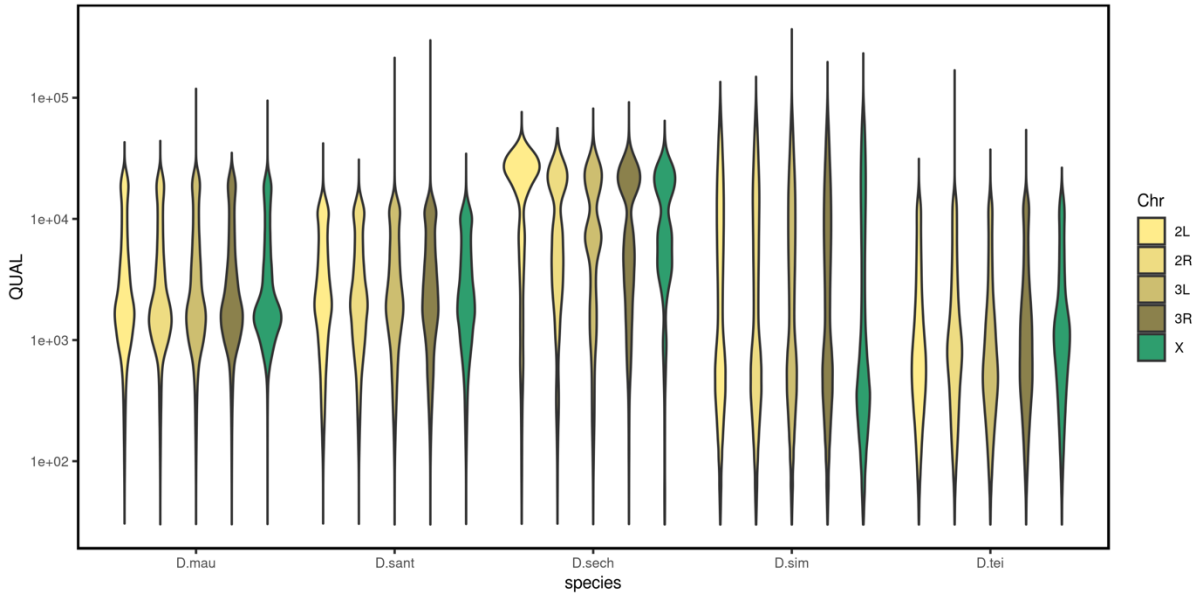

Figure S8. Phred-scaled quality score across species. The distribution is similar across autosomes and the X chromosome for *D. mauritiana, D. santomea. D. sechellia, D. simulans* and *D. teissieri*. The y-axis is log-scaled for better visualization.

Text S1: The effect of background selection on diversity

To estimate the impact of BGS on diversity levels relative to a neutral or near neutral genomic region, we followed the calculations by Nam et al. 2015 based on previous theory (Durret 2008; Nam et al. 2015). BGS reduces diversity locally by purging deleterious mutations accompanied by linked neutral variation. The strength of BGS is determined by the deleterious mutation rate $U=\mu L$, the recombination rate $R=\rho L$, and the selection strength of the deleterious mutations s_d. μ is the per locus deleterious mutation rate, ρ is the per locus recombination rate, and L is the length of the region under selection.

A higher deleterious mutation rate in a given region of the genome is expected to lead to stronger BGS as more linked neutral variation will be removed from the population due to selection on deleterious variants. In regions with a high recombination rate, BGS is expected to be weak as recombination breaks down haplotypes and reduces the size of the region affected by selection. Finally, when selection is weak but not effectively neutral ($s > 1/Ne$) haplotypes containing the deleterious alleles will have the chance of rising in frequency before they are purged from the population leading to stronger BGS. This means that a region with a higher deleterious mutation rate, low recombination and/or weak selection will show lower levels of diversity due to stronger BGS.

Here, we compare the expected reduction in π due to BGS between a window subject to strong BGS (greater expected reduction in π) and a window subject to weak BGS (less expected reduction in π). To do so we define the factors of f_U, f_R and f_s as in Nam et al 2015. These parameters represent the relative reduction in mutation rate, recombination rate and selection strength in low versus high diversity regions. The expected relative reduction in low diversity (π_{low}) windows relative to high diversity windows (π_{high}) is given by:

$$\frac{\pi_{low}}{\pi_{high}} = \frac{\exp\left(-\frac{f_U U}{f_s s_d + f_R R}\right)}{\exp\left(-\frac{U}{s_d + R}\right)}$$

D. sechellia is the species with the lowest $Ne \sim 1\text{x}10^5$ (Kliman et al. 2000; Legrand et al. 2009) in our study, while the rest of the species are on the order of $Ne \sim 1\text{x}10^6$. From this, we take $s_d = 1\text{x}10^{-5}$, $1\text{x}10^{-4}$ and $1\text{x}10^{-3}$ which correspond to an Nes_d between 1 and 100 in the lower $Ne=$

$1x10^5$ case and a Nes_d between 10 and 1000 in the Ne= $1x10^6$ case. As explained above, when s_d is weak the effect on diversity is stronger. Therefore, we tested three f_s values f_s=0.1, 0.5 and 1 so that selection at low diversity windows was $0.1s_d$,$0.5s_d$ and equal to s_d, where s_d is the deleterious selection strength outside low diversity windows.

We use the Comeron et al. (Comeron et al. 2012) recombination map from *D. melanogaster* to compute an estimate of f_R from the low and high diversity regions in our two populations of *D. melanogaster.* For the Zambia population we obtain $f_R = 0.44$ while for the Raleigh population we obtain $f_R = 0.59$, so we use the average $f_R = 0.5$ for our calculations.

Mutation rates in *Drosophila* are on the order of 10^{-9} mutations per base pair. We tested different three different proportions of deleterious mutation rate where 10%, 50% or 80% or the incoming mutations were considered to be deleterious. For each of these, we varied the relative mutation rate in low versus high diversity regions (f_U) using f_U = 1, 5 and 10. An f_U = 1 means that the deleterious mutation rate is equal across windows while an f_U = 10 means that the deleterious mutation rate is 10 times higher in the window subject to strong BGS.

With the parameters as described above, we computed the expected π_{low}/π_{high}. **Figure S4** shows that BGS can reduce diversity below 20% of the expectation under neutrality only in the most extreme cases, in which 80% of the mutations in the region are deleterious and mutation rate is >10 times higher in low diversity windows.

Text S2: Overlapping orthologous gene groups on the X chromosome across species

To investigate whether the low diversity regions observed across species on the X chromosome correspond to similar functions under selection across species, we identified orthologous gene groups present in two or more species for regions both below and above the low diversity threshold (**Methods**). We found that, excluding *D. melanogaster* (ZI) versus *D. melanogaster* (RA) comparisons, the proportion of orthologous genes that are present across *n*=2,3 or 4 species was significantly higher in low diversity regions on the X compared to the rest of the chromosome (**Fig. S5**). Specifically, we observed that 10%, 3% and 0.5% of orthologs in low diversity windows were shared across 2, 3 or 4 species, respectively. In contrast, for regions above

the threshold, only 7%, 0.5% and 0.01% of orthologs where identical across 2, 3 or 4 species, respectively. Moreover, when we included all populations, we found the highest overlap was between *D. melanogaster* (ZI) and *D. melanogaster* (RA) (14/ 43 orthologous gene groups found in at least two species; see **Fig. S6, Fig. S7**). This finding aligns with our expectations since *D. melanogaster* (ZI) and *D. melanogaster* (RA) are two populations of the same species, suggesting that the selection process may have originated in the ancestral population and persisted in the derived population. Despite the several instances of shared targets of selection across species, the majority of genes found in low-diversity regions were found in a single species (**Fig S5, Fig. S6**), indicating that distinct genes are targeted by selection independently.

References

Comeron JM, Ratnappan R, Bailin S. 2012. The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genet* 8.

Durret R. 2008. Probability Models for DNA Sequence Evolution. Springer, Berlin

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the Drosophila simulans complex species. *Genetics* 156.

Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou ML. 2009. Species-wide genetic variation and demographic history of Drosophila sechellia, a species lacking population structure. *Genetics* 182.

Nam K, Munch K, Hobolth A, Dutheil JY, Veeramah KR, Woerner AE, Hammer MF, Prado-Martinez J, Sudmant PH, Kidd JM, et al. 2015. Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc Natl Acad Sci U S A* 112:6413–6418.