Neutral sequence evolution in short introns of *Drosophila melanogaster*

Burçin Yıldırım[a,b], Claus Vogl[a,b,*]

[a] *Department of Biomedical Sciences, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*
[b] *Vienna Graduate School of Population Genetics, A-1210 Wien, Austria*

Acknowledgements

The authors wish to thank all members of the Vienna Graduate School of Population Genetics for support, and Lynette Caitlin Mikula for critically reading the article and providing helpful suggestions. This work was supported by the Austrian Science Fund (FWF; W1225-B20).

Competing interests

The authors declare no competing interests.

*Corresponding author
Email addresses: burcin.yildirim@vetmeduni.ac.at (Burçin Yıldırım), claus.vogl@vetmeduni.ac.at (Claus Vogl)

September 4, 2023

Abstract

Population genetic inference of selection on the nucleotide sequence level often proceeds by comparison to a reference sequence evolving only under mutation and population demography. Among the few candidates for such a reference sequence is the 5' part of short introns (5SI) in *Drosophila*. In addition to mutation and population demography, however, there is evidence for a weak force favoring *GC* bases, likely due to *GC*-biased gene conversion (gBGC), and for the effect of linked selection. Here, we use polymorphism and divergence data of *Drosophila melanogaster* to detect and describe the forces affecting the evolution of the 5SI. We separately analyze mutation classes, compare them between chromosomes, and relate them to recombination rate frequencies. *GC*-conservative mutations seem to be mainly influenced by mutation and drift, with linked selection mostly causing differences between the central and the peripheral (*i.e.,* telomeric and centromeric) regions of the chromosome arms. Comparing *GC*-conservative mutation patterns between autosomes and the X chromosome, showed differences in mutation rates, rather than linked selection, in the central chromosomal regions after accounting for differences in effective populations sizes. On the other hand, *GC*-changing mutations show asymmetric site frequency spectra, indicating the presence of gBGC, varying among mutation classes and in intensity along chromosomes, but approximately equal in strength in autosomes and the X chromosome.

Keywords: neutral evolution, *GC*-biased gene conversion, recombination, linked selection, mutation rate, X-chromosome, *Drosophila*

Introduction

DNA sequence polymorphism and divergence data have been used to infer evolutionary processes in many fields of evolutionary biology, from molecular evolution to anthropology. Whether the aim is phylogenetic reconstruction, demographic inference, or detection of selection, it is difficult to tease apart the different population genetic forces that determine DNA sequence patterns. According to the neutral theory of molecular evolution (Kimura, 1983), mutations can be either selectively neutral, subject only to mutation and random drift, or strongly selected, leading to rapid fixation or elimination. Early analysis methods often tested for deviation from neutral equilibrium (*e.g.,* Tajima, 1989; Fu & Li, 1993). However, with the availability of genome-wide data and advanced analytical methods, it has become clear that only very rarely sequences evolve neutrally, free from the effect of directional or more complicated selective forces, and are at equilibrium. Furthermore, the indirect effect of selection on linked neutral sites (background selection or selective sweeps) on molecular diversity patterns has been shown (Charlesworth et al., 1993; Schrider et al., 2016). The relative contributions of the different population genetic forces may vary depending on the organism under study and among different genomic regions. Ignoring a particular evolutionary force or its interplay with other forces may introduce biases into population genetic inferences (Lartillot, 2012; Bolívar et al., 2016, 2018; Galtier et al., 2018; Borges et al., 2019; Bolívar et al., 2019; Boman et al., 2021).

Fourfold degenerate sites, or generally synonymous sites, have been regarded as neutrally evolving, because an exchange of a nucleotide does not affect the encoded amino acid. Even though all genomes show some degree of codon usage bias in these sites (Hershberg & Petrov, 2008), the intensity of selection has been generally considered weak in relation to drift. Therefore, such sites have been used as a neutral reference for many tests (*e.g.,* McDonald & Kreitman, 1991; Yang et al., 2000). These are generally based on the infinite sites model and assume reduction in polymorphism and divergence due to purging of deleterious mutant alleles with directional selection. However when selection is weak and opposes the mutation bias, it can increase polymorphism and divergence levels compared to neutrality (McVean & Charlesworth, 1999; Vogl & Mikula, 2021), creating unexpected patterns. An additional challenge for using fourfold degenerate sites as a neutral reference comes from studies showing that codon usage bias may vary from very weak to strong depending on the species, the codon and the position in the genome (Chamary et al., 2006; Lawrie et al., 2013). Thus, it is not clear *a priori* which codon is favored for a specific site in a specific species, and preferences may evolve.

A similarly weak force is GC-biased gene conversion (gBGC), which is tied to the repair of double-strand breaks (DSBs) by homologous recombination. When a site is heterozygous for a strong (S; G and C bases) and a weak (W; A and T bases) allele, the heteroduplex mismatch formed during the DSB repair might resolve biased towards the strong base (Marais, 2003). The biased resolution of these GC-changing mutations ($S \leftrightarrow W$) leads to a nonadaptive directional force, indistinguishable in its effect from selection for GC bases (Duret & Galtier, 2009; Bolívar et al., 2018). At the molecular level, the repair efficiency of GC-changing mutations might change between transitions and transversions (Dohet et al., 1985; Holmes et al., 1990), and the effects of it might also be observed in the gBGC dynamics at the population genetic level (Lartillot, 2012; Bergman & Schierup, 2021). Studies have shown that failing to account for gBGC may lead to biased inference of selection and demography (Bolívar et al., 2018; Pouyet et al., 2018; Galtier et al., 2018). Additionally, the effects of gBGC on diversity may be falsely interpreted as a consequence of linked selection or Hill-Robertson interference (Bolívar et al., 2016; Boman et al., 2021) and confound the

detection of expectations from nearly neutral theory (Bolívar et al., 2019) due to its relationship with recombination.

In flies of the genus *Drosophila*, it is known that a large fraction of the genome, including intronic and intergenic noncoding sequences, is under selective constraints (Andolfatto, 2005; Haddrill & Charlesworth, 2008). Synonymous codons are also subject to selection due to codon usage bias (Akashi, 1994, 1995) with variable intensity depending on species and codons (Singh et al., 2009; Zeng, 2010). So far, the best candidates for neutrally evolving, unconstrained sites in *Drosophila* are the nucleotides at position 8-30 on the 5' end of introns shorter than 65 bp (hereafter 5SI) (Halligan & Keightley, 2006; Parsch et al., 2010; Clemente & Vogl, 2012; Yıldırım & Vogl, 2023). These sites exhibit higher divergence and polymorphism levels compared to other regions in introns (Parsch et al., 2010). Longer introns contain more functional elements, as shown by a negative correlation between divergence and length (Haddrill et al., 2005) and most other sequences inside short introns are likely under selection due to their association with splicing (Yıldırım & Vogl, 2023). Thus, many studies utilized this short intronic region to infer directional selection on fourfold degenerate sites (Lawrie et al., 2013; Machado et al., 2020) or demography before detecting sweep signatures (Garud et al., 2015).

Despite the utility of the 5SI as a neutral reference in population genetic analyses, biallelic frequency spectra of weak vs strong bases deviate from the neutral prediction of symmetry, with an excess of high-frequency GC variants (Clemente & Vogl, 2012; Jackson et al., 2017; Jackson & Charlesworth, 2021), which indicates the presence of a directional force, likely gBGC. The presence of gBGC in the fruit fly *Drosophila* has been long debated and remained inconclusive. Clemente & Vogl (2012) explained the asymmetry in site-frequency spectra (SFS) of AT-to-GC polymorphism in *D. melanogaster* by a shift in mutation bias towards AT and a context-dependent mutational pattern. Robinson et al. (2014) claimed that the effect of gBGC is unlikely to impact genome evolution patterns. Other studies showed evidence for gBGC, but differed in their claims on which chromosome and in which species it operates (Haddrill & Charlesworth, 2008; de Procé et al., 2012; Jackson et al., 2017).

Most recently, Jackson & Charlesworth (2021) demonstrated the existence of a GC-favoring directional force in autosomal 5SI sites of both *D. simulans* and *D. melanogaster*, using unfolded SFS of GC-changing mutations ($S \to W$, $W \to S$) from a larger population data set and better reference genomes than the previous studies. This immediately raises interesting research questions: Does gBGC affect transitions and transversions similarly? Is gBGC also present in the X chromosome of *Drosophila*? If the pattern of gBGC is recombination-dependent, we expect about equal effects on autosomes and the X chromosome in *Drosophila*, as there is no recombination in males. Do we therefore find similar effects of gBGC on the X chromosome? Most importantly, it is known from other species that the presence of gBGC can lead to biased interpretations of the effects of linked selection and other nonadaptive forces. Variation in polymorphism along chromosomes that has been shown to be correlated with recombination rate variation in *Drosophila* is accepted to be the result of linked selection, but may also be affected by gBGC. How does the previously neglected presence of gBGC in *Drosophila* influence our interpretation of neutral sequence evolution patterns in autosomes and the X chromosome? Are patterns that have been attributed to linked selection affected or actually caused by gBGC?

In this study, we address these questions by utilizing 5SI from one of the largest and most accurate polymorphism data from the ancestral population of *D. melanogaster* and divergence data with *D. simulans* (Rogers et al., 2014; Lack et al., 2015; Jackson et al., 2017). The divergence time between these two species is conveniently so small that very few double mutations separating

the species are expected, but large enough that little shared polymorphism is expected (estimates are given in the Results section). Therefore we can assume that only single mutations give rise to polymorphic and divergent sites. This allows us to compare estimates of polymorphism and divergence among different mutation classes and between autosomes and the X chromosome to identify and describe the relative contributions of various population genetic forces, such as gBGC, mutation, and drift, in shaping the evolution of this presumably neutral sequence class. Specifically, we contrast GC-conservative mutation classes, which are expected to be unaffected by gBGC, with GC-changing mutations to answer the following questions: What is the relative effect of linked selection and gBGC on the variation in polymorphism and divergence i) along chromosomes and ii) between autosomes and X chromosomes? iii) How does variation in recombination rates modulate these effects? While addressing these questions, we also provide a detailed description of the dynamics of gBGC separately for transitions and transversions, both along and between chromosomes.

Materials and Methods

Data Used in the Analyses

We analyzed previously published whole-genome data from a population of *D. melanogaster* from the ancestral range in Zambia (Lack et al., 2015). After excluding individuals showing admixture with European populations (Lack et al., 2015), the dataset consists of 69 individuals both for autosomes and X chromosome. Sequences were obtained as consensus FASTA files, and the full description of the data processing (sampling, sequencing, variant calling) can be found in Lack et al. (2015). We also obtained consensus FASTA files from a population sample of *Drosophila simulans* from Madagascar including 21 individuals for all chromosomes (Rogers et al., 2014; Jackson et al., 2017).

Using annotations from the reference genomes of *D. melanogaster* (r5.57 from *http://www.flybase.org/*) and *D. simulans* (Hu et al., 2013), orthologous intron coordinates were extracted and alignments of all samples were created. To avoid including the same intron sequence more than once due to alternatively spliced isoforms in the GFF file (see *https://www.ensembl.org/info/website/upload/gff.html*; last accessed November 1, 2020), only one entry of introns with overlapping coordinates were used. Bases in position 8-30 in short introns ($\leq$ 65 bp) were extracted to use as a proxy of least constrained sites (Halligan & Keightley, 2006; Parsch et al., 2010).

Polymorphism and divergence estimates

We inferred site frequency spectra from the Zambian *D. melanogaster* samples for all six possible combinations of base pairs, for both autosomal and X-linked short introns. We filtered out sites that overlapped coding sequences, contained an undefined nucleotide state in at least one of the sequences in the sample alignment and sites with more than two alleles. To estimate scaled mutation rates $\theta_{ij} = 4N_e\mu_{ij}$, let i and j stand for the four bases $i, j \,\epsilon\, (A, T, G, C)$. We assumed that mutation rates are low, such that more than two alleles can not segregate in the population, and a strand symmetric rate matrix, where nucleotides, A and T or G and C are interchangeable. This allows the use of the ML estimators from Vogl et al. (2020). With SFS data available from L loci with M genomes, let L_{ij} be the counts of polymorphic sites and L_i the counts of monomorphic

sites. The six maximum likelihood estimators (MLEs) for the scaled mutation rates then are:

$$
\begin{aligned}
(\theta_{AT}, \theta_{TA}): \quad & a_p = \frac{L_{AT}}{L^{(AT)} + 1/2L^{(AT,CG)}} \frac{1}{H_{M-1}} \\
(\theta_{GC}, \theta_{CG}): \quad & f_p = \frac{L_{CG}}{L^{(CG)} + 1/2L^{(AT,CG)}} \frac{1}{H_{M-1}} \\
(\theta_{CT}, \theta_{GA}): \quad & b_p = \frac{L^{(AT,CG)}}{L^{(CG)} + 1/2L^{(AT,CG)}} \frac{b'}{2H_{M-1}} \\
(\theta_{TC}, \theta_{AG}): \quad & c_p = \frac{L^{(AT,CG)}}{L^{(AT)} + 1/2L^{(AT,CG)}} \frac{1-e'}{2H_{M-1}} \\
(\theta_{CA}, \theta_{GT}): \quad & d_p = \frac{L^{(AT,CG)}}{L^{(CG)} + 1/2L^{(AT,CG)}} \frac{1-b'}{2H_{M-1}} \\
(\theta_{AC}, \theta_{TG}): \quad & e_p = \frac{L^{(AT,CG)}}{L^{(AT)} + 1/2L^{(AT,CG)}} \frac{e'}{2H_{M-1}} \,,
\end{aligned}
\tag{1}
$$

where $L^{(AT)} = L_A + L_T + L_{AT}$, $L^{(CG)} = L_C + L_G + L_{CG}$, $L^{(AT,CG)} = L_{AC} + L_{AG} + L_{TC} + L_{TG}$ and $H_{M-1} = \sum_{y=1}^{M-1} = 1/y$. The parameters b'_p and e'_p correspond to $\frac{b}{b+d}$ and $\frac{e}{e+c}$, respectively. They are obtained by maximizing the following log likelihood:

$$
\log L(e'_p, b'_p) = \sum_{y=1}^{M-1} \left\{ (l_{AC}(y) + l_{TG}(y)) \log\left(\frac{e'_p}{M-y} + \frac{1-b'_p}{y} \right) + (l_{AG}(y) + l_{TC}(y)) \log\left(\frac{1-e'_p}{M-y} + \frac{b'_p}{y} \right) \right\}. \tag{2}
$$

From polymorphism data, we also calculated mutation bias towards AT (β) as follows;

$$
\beta = \frac{L^{(AT)} + 1/2L^{(AT,CG)}}{L} \,. \tag{3}
$$

As a measure of divergence between *D. melanogaster* and *D. simulans*, we used an estimate corresponding to D_a, the net nucleotide differences between populations since the species split (Nei & Li, 1979). It is defined as $D_a = D_{xy} - (\pi_x + \pi_y)/2$, where the expectation of the pairwise difference D_{xy} is $2\mu t + \theta_{anc}$. D_a uses the average current levels of polymorphisms ($(\pi_x + \pi_y)/2$) as a measure of ancestral polymorphism (θ_{anc}), and subtracts this value from the total divergence to get an estimate of $\delta = 2\mu t$, where t is the time since the species split. If linked selection was present before the split of the two populations, in the ancestral population, D_a should be affected only slightly, while D_{xy} should be reduced (figure 1). Begun et al. (2007) showed that current targets of selection in *Drosophila* are also targets of recurrent selection, thus using a D_a-like measure for divergence would minimize the effect of linked selection. We calculated such a divergence estimate from the alignment of two *D. melanogaster* and two *D. simulans* samples. Corresponding to the estimators

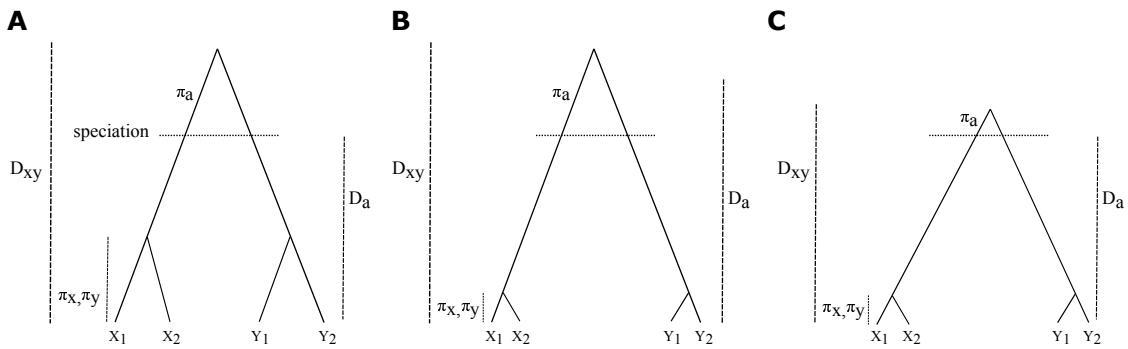

Figure 1: Genealogies of four samples (X_1, X_2, Y_1, Y_2) from two populations or species (X and Y) and measures of divergence with (B, C) and without (A) the effect of linked selection. π_x, π_y and π_a represent the polymorphism level in current (X, Y) and ancestral populations, respectively. D_{xy} is a measure of divergence defined as average number of pairwise differences between sequences of populations. Lastly, D_a is a measure of pairwise differences since the split of populations. Panel A represents a scenario without linked selection, panel B represents the linked selection affecting current-day populations and panel C shows a scenario with recurrent linked selection affecting both ancestral and extant populations. Figures are adapted from Cruickshank & Hahn (2014).

of scaled mutation rates, six estimators of divergence were obtained for pairs of $\delta_{ij} = 2\mu_{ij}t$:

$$
\begin{aligned}
(\delta_{AT}, \delta_{TA}): \quad & a_d = \frac{D_{AT} - L_{S,AT}}{L^{AT} + 1/2L^{(AT,CG)}} \\
(\delta_{GC}, \delta_{CG}): \quad & f_d = \frac{D_{CG} - L_{S,CG}}{L^{CG} + 1/2L^{(AT,CG)}} \\
(\delta_{CT}, \delta_{GA}): \quad & b_d = \frac{D_{AG} + D_{TC} - L_{S,AG} - L_{S,TC}}{L^{CG} + 1/2L^{(AT,CG)}} \\
(\delta_{TC}, \delta_{AG}): \quad & c_d = \frac{D_{AG} + D_{TC} - L_{S,AG} - L_{S,TC}}{L^{AT} + 1/2L^{(AT,CG)}} \\
(\delta_{CA}, \delta_{GT}): \quad & d_d = \frac{D_{AC} + D_{TG} - L_{S,AC} - L_{S,TG}}{L^{CG} + 1/2L^{(AT,CG)}} \\
(\delta_{AC}, \delta_{TG}): \quad & e_d = \frac{D_{AC} + D_{TG} - L_{S,AC} - L_{S,TG}}{L^{AT} + 1/2L^{ATCG}} \,,
\end{aligned}
\tag{4}
$$

where D_{ij} corresponds to the number of sites differentially fixed for i and j nucleotides and $L_{S,ij}$ to the number of ancestral shared polymorphism segregating for i and j in *D.melanogaster* and *D. simulans*.

Under equilibrium conditions, we note that for autosomes the expected neutral divergence $\mathrm{E}[a_d]$ and the expected neutral scaled mutation rate or polymorphism $\mathrm{E}[a_p]$ are proportionally

affected by the same pair of mutation rates μ_{AT} and μ_{TA}, such that we have:

$$\begin{aligned}\frac{t}{2N} &= \frac{\mathrm{E}[a_d]}{\mathrm{E}[a_p]} \\ &\approx \frac{D_{AT} - L_{AT}/H_{M-1}}{L^{AT} + 1/2L^{(AT,CG)}} \frac{L^{AT} + 1/2L^{(AT,CG)}}{L_{AT}/H_{M-1}} \\ &= \frac{D_{AT} - L_{AT}/H_{M-1}}{L_{AT}/H_{M-1}} = \frac{D_{AT}H_{M-1}}{L_{AT}} - 1\,,\end{aligned} \tag{5}$$

and similar for other pairs of polymorphism and divergence estimators. Assuming equality of male and female effective population sizes and accounting for the hemizygosity of males, the ratio of divergence over polymorphism for the autosomes would be 3/4 of the X chromosome:

$$\frac{t}{2N} = \frac{\mathrm{E}[a_d]}{\mathrm{E}[a_p]} = \frac{3}{4}\frac{\mathrm{E}[a_d^{(X)}]}{\mathrm{E}[a_p^{(X)}]}\,. \tag{6}$$

When a mutation class without a directional force is compared to another with a directional force γ, the "neutrality index" can be calculated as ratio of ratios of divergence and polymorphism between the two classes assuming neutrality. Importantly, the mutational terms cancel when the ratio is taken. Let r_N stand for the ratio of expected divergence to expected polymorphism of the neutral class and r_γ for that with a directional force. Then $r_N/r_\gamma = ((e^\gamma/2 - e^{-\gamma/2})/\gamma)^2$ (see equation 54 in Vogl & Mikula (2021)), which is always greater than 1 with $\gamma \neq 0$ in equilibrium.

Due to filtering out the non-biallelic sites and missing polymorphism in *D. melanogaster* population data, different numbers of sites were available for polymorphism and divergence based analyses. To make analyses comparable, we created a reduced dataset that included only the common sites, of which 137,699 were autosomal and 16,873 X chromosomeal. To compare, we also analyzed polymorphism and divergence data from *D. simulans* similarly, resulting in 150,613 autosomal and 19,412 X chromosomal common sites. The 95% confidence intervals for each point estimate were determined from 1000 bootstrap-resamples. Our estimates were calculated for whole chromosomes, as well as for the central and peripheral (telomeres and centromeres combined) regions of chromosome arms. Unless otherwise stated, our reported estimates come from the central part of the chromosome arms after excluding telomeres and centromeres.

Recombination rates and base composition

To study the co-variation of the inferred parameter estimates and recombination rates, we retrieved the recombination rate estimates, based on crossover events, from the *D. melanogaster* recombination map of Comeron et al. (2012). We divided the dataset into bins with approximately equal number of observations, before and after excluding peripheral regions of the chromosome arms with low recombination rates. Mean recombination rate for each bin is given in table S1. The site frequency spectra inferred separately for each recombination bin were used to estimate the polymorphism, divergence and the strength of gBGC.

We examined the base composition by looking at *GC* content, calculated for each intron separately from the reference sequences of *D. melanogaster* and *D. simulans*. The *GC* content of the four-fold degenerate sites (FFDS) from the same genes that introns located were also obtained and used as a proxy for background base composition. To investigate the variation of *GC*-biased gene conversion with the base composition, we divided the dataset into five bins with approximately

equal number of observations depending on the background GC content (*i.e.,* FFDS GC content). The range of GC content for each bin is given in tables S2 and S3 for *D. melanogaster* and *D. simulans*, respectively.

Estimates of gBGC strength

Heteroduplex mismatches formed during the repair of double strand breaks can involve either pairing between the bases G and C (strong: S:S), A and T (weak: W:W) or between strong and weak bases, S:W. Preferential resolution of the S:W mismatches into $G:C$ rather than $A:T$ leads to GC-biased gene conversion (gBGC) (Marais, 2003). Since, gBGC only affects S:W mismatches they are referred to as GC-changing, while the others (S:S, W:W) are called GC-conservative. This categorization allows us to use unpolarized data to estimate GC bias while considering the site frequency spectra (SFS) of all six possible nucleotide pairs (Borges et al., 2019). gBGC is a directional force quantified by $B = 4N_e b$, b is the conversion bias that depends on recombination rate, tract length and repair bias towards GC (Nagylaki, 1983).

We inferred the strength of this directional force, gBGC, under mutation-drift-directional force equilibrium by using the maximum likelihood estimator of Vogl & Bergman (2015). Assuming low mutation rate and binomial sampling, the probability of a mutation segregating at frequency y in the limit of large M is;

$$\Pr(y \,|\, \vartheta, B, M) = \vartheta e^{By/M} \frac{M}{y(M-y)}\,, \tag{7}$$

where $\vartheta = ((1-\beta)\beta\theta)/((1-\beta)e^B + \beta)$ and β is mutation bias towards AT. The likelihood of the polymorphic loci is sufficient for the inference of B and expressed as:

$$\Pr(L_1, ..., L_{M-1} \,|\, B, M) = \prod_{i=1}^{M-1} \left(\frac{e^{By/M} \frac{M}{y(M-y)}}{\sum_{i=1}^{M-1} e^{By/M} \frac{M}{y(M-y)}} \right)^{L_y}, \tag{8}$$

where L_y is the number of sites with y GC-changing mutations. B estimates were obtained by maximizing the likelihood in equation 8 for the SFS from GC-changing polymorphisms (A/C, A/G, T/C and T/G) of the 5SI. The SFS from GC-conservative polymorphisms (A/T, G/C) was considered as putatively neutral control ($B = 0$). We performed likelihood-ratio tests (LRT) to compare between the different nested models. Conditional on B, we also estimated the mutation bias towards AT ($\hat{\beta}$). For the inference, the mutation bias parameter was set to $\varrho = 1 - ((1-\beta)e^B)/((1-\beta)e^B + \beta)$ with the maximum likelihood estimate of;

$$\hat{\varrho} = 1 - \frac{L_M}{L} + \frac{L_p e^B \sum_{i=1}^{M-1} e^{-By/M} \frac{M}{y(M-y)}}{L(\sum_{i=1}^{M-1} e^{By/M} \frac{M}{y(M-y)} + e^B \sum_{i=1}^{M-1} e^{-By/M} \frac{M}{y(M-y)})}\,. \tag{9}$$

Given the estimate of $\hat{\varrho}$, $\hat{\beta}$ was recovered using $1 - \frac{\hat{\varrho}}{\hat{\varrho} + e^B(1-\hat{\varrho})}$.

We also inferred B while correcting for the effect of demography and population structure by introducing noise parameters r_y to the likelihood (equation 8)(Bergman & Schierup, 2021). We obtained r_y by comparing the neutral expectation of the SFS with the empirical SFS of the neutral sites, *i.e.,* SFS of GC-conservative polymorphisms.

$$r_y = \left(\frac{L_{y,n}}{\sum_{i=1}^{M-1} L_{y,n}} \right) \Big/ \left(\frac{1/(y(M-y))}{\sum_{i=1}^{M-1} 1/(y(M-y))} \right), \tag{10}$$

where $L_{y,n}$ is the number of sites with y GC-conservative polymorphisms. Due to the large sample sizes M and splitting of the data according to mutation classes and genomic regions, counts of $r_y = 0$ are possible, which leads to undefined values when calculating the log-likelihood. To avoid that, we added pseudo-counts proportional to the equilibrium expectations $1/y + 1/(M - y)$. We note that accounting for demography in this way lowers the statistical power.

Results

We investigated the pattern of molecular evolution and variation in presumably neutral short introns of *Drosophila melanogaster* and *D. simulans* by comparing different nucleotide classes along chromosomes, between chromosomes and in relation to recombination rates and GC-content to estimate the contribution of different forces to the observed patterns. We used both polymorphism and divergence data to get estimates of scaled mutation rates (or polymorphism or diversity, $4N_e\mu_{ij}$) and divergence ($2\mu_{ij}t$) for each nucleotide class. Instead of a bi-allelic mutation-drift model, we used a multi-allelic model, corresponding to the four bases, that can provide information about the possible differences in mutational bias and rate between different alleles. Given the four bases, this would imply $4 \cdot 3 = 12$ parameters.

If there is no DNA strand-specificity of mutation rates, the equal proportion of complementary bases (*i.e.,* of the weak bases A and T and the strong bases C and G, respectively) along the DNA leads to strand symmetry, *i.e.,* Chargaff's second parity rule (Mitchell & Bridge, 2006). The presumably unselected intronic sequences we analyzed exhibit only very minor deviations from strand symmetry (slight biases towards T and C over A and G, respectively) (Bergman et al., 2017), which are attributed to neutral processes, such as transcription-coupled asymmetries (Touchon et al., 2004). Thus, we obtained the MLE of scaled mutation rates by assuming neutral equilibrium and strand-symmetric mutation (Vogl et al., 2020), which reduced the number of parameters from twelve to six $(a_p, b_p, c_p, d_p, e_p, f_p)$, see equation (1). Our inference further assumes that the scaled mutation rates are small, ≤ 0.025 (Vogl & Clemente, 2012), which indeed holds true (table 1). Therefore segregation of more than two alleles in the sample is negligible.

In addition to mutational differences, some of these nucleotide classes might be affected distinctly by other evolutionary forces. The $A : T$ (a_p) and $C : G$ (f_p) polymorphisms are transversions that exhibit negligible mutation bias in *Drosophila* and, more importantly, these two mutation classes are GC-conservative, meaning that they are unaffected by gBGC. The other four mutation classes are GC-changing and include both transitions (b_p, c_p) and transversions (d_p, e_p). They are affected by gBGC; as stated before, transition and transversion mutations may be differently affected by gBGC (Lartillot, 2012; Bergman & Schierup, 2021). This justifies going beyond the usual classification of GC-changing (S $\leftrightarrow$ W) and GC-conservative (S $\leftrightarrow$ S or W $\leftrightarrow$ W) mutations (Bolívar et al., 2016; Boman et al., 2021).

We inferred the divergence with a D_a-like measure for six nucleotide classes $(a_d, b_d, c_d, d_d, e_d, f_d)$ as above (see equation 4). Divergence is influenced little by demography, *i.e.,* changes in effective population size, as fluctuations in population size converge to the harmonic mean over typical divergence times. But while divergence is little affected by demography, it is affected by gBGC and directional selection (*e.g.,* Kimura, 1962; Vogl & Mikula, 2021). As long as divergence times are relatively small, such that double mutations are too rare to influence inference, the independence of a_p and f_p polymorphisms from gBGC also extends to $A : T$ and $C : G$ divergence. Additionally, when divergence times are large enough, little shared heterozygosity is expected. With these conditions met, divergence and polymorphism ratios can be compared between nucleotide classes and

chromosomal regions to disentangle the effect of population genetic forces.

Generally, the expected shared heterozygosity between two populations decreases at a rate of t/N, such that after $t = N$ generations, only 0.37 of the polymorphism is shared. In the case of *D. melanogaster* and *D. simulans*, we estimate $t/N \approx 4$ (see table 2). Therefore, the proportion of heterozygosity shared between the species is expected to be less than 0.02. Furthermore the probability of observing double mutations in these two populations is approximately $(2\mu t)^2$. An estimate of $2\mu t$ can be obtained by multiplying the estimated divergence t/N by the estimated expected heterozygosity $\theta = 4\mu N$ within a population. For *D. melanogaster*, $\theta = 4\mu N$ is about 0.024 (table 1). Consequently, we expect about 0.002 double mutations between *D. melanogaster* and *D. simulans*. We therefore treat shared polymorphism between the two species and double mutations as negligible and the twelve parameters (six mutation classes for scaled mutation rates and divergence, respectively) as independent. In the following, we will drop the subscripts when we refer to the mutation classes, as it is clear from the context whether polymorphism or divergence or ratios of both are implicated.

Diversity estimates along chromosomes

Our estimates of the overall expected heterozygosity (see equation 81 Vogl et al., 2020) at 5SI sites without differentiating the nucleotides are identical to the results of previous studies (Parsch et al., 2010; Jackson & Charlesworth, 2021): X chromosomal heterozygosity is slightly higher or equal to the autosomal heterozygosity and for all chromosomes estimates are lower towards the peripheral regions (telomeres and centromeres) of chromosome arms (table 1). The latter is expected as in the *Drosophila* genome recombination rates are lower towards telomeres and centromeres (Comeron et al., 2012) and nucleotide polymorphism is reduced in regions with reduced recombination (Begun & Aquadro, 1992). It is also known that mutations are AT biased in *Drosophila* (Vogl & Bergman, 2015) and accordingly our estimates of mutation rates going from G or C to A or T (b, d) is higher than in their reverse direction (c and e, respectively) (figure 2). The degree of this bias is lower for the X chromosome ($\beta_A = 0.668$ (95% CI: $0.664 - 0.671$) and $\beta_X = 0.621$ (95% CI: $0.613 - 0.630$)). Additionally, estimates are higher for transitions (b, c) than for transversions (a, f, d, e) with a ratio of 2.18 (95% CI: $2.09 - 2.27$). Note that the estimate of b, both a GC-to-AT mutation and a transition, is the highest.

Table 1: The overall expected heterozygosity estimates for different parts of the autosomes and X-chromosome (95% CIs in brackets). WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

	Autosome	X
WChr	$0.0196\ (0.0188, 0.0204)$	$0.0208\ (0.0186, 0.0231)$
CChr	$0.0239\ (0.0229, 0.0249)$	$0.0238\ (0.0212, 0.0266)$
PChr	$0.0099\ (0.0089, 0.0109)$	$0.0103\ (0.0071, 0.0138)$

The relative relationship between the nucleotide classes for the divergence estimates mirrors the patterns in scaled mutation rates. However, contrary to scaled mutation rates, divergence does not decrease towards the periphery, instead there is a slight increase for some GC-changing mutations in autosomes. This shows that mutation rate variation does not greatly contribute to a positive correlation between recombination and polymorphism, rather the variation in the effective population size N_e due to the direct or indirect effect of directional forces must cause it. If the

major driver of the reduction in polymorphism levels is linked selection, we do not expect differences between mutation classes, while a directional force, like GC-biased gene conversion (gBGC), might differentially affect mutation classes. The presence of gBGC should be apparent from the differences between GC-changing (b, c, d, e) and GC-conservative (a, f) mutations, which we indeed observe. However, comparing directly polymorphism or divergence estimates between mutation classes is not helpful: as GC-changing mutations include both transitions and transversions but GC-conservative ones only transversions, we still would not be able to distinguish between the effects of mutational and directional (gBGC) forces. Thus, for each mutation class, we get the ratio of divergence over polymorphism, which is minimally affected by mutation and is expected to scale inversely with the effective population size.

For all chromosomes and chromosomal parts, the divergence over polymorphism ratio of GC-changing mutations is significantly higher compared to GC-conservative ones (table 2). This shows that linked selection should not be the only driver creating nucleotide polymorphism variation along the genome and supports the presence of a directional force differing between these two classes of mutations, likely gBGC. But a directional force differing between them should result in a lower divergence to polymorphism ratio in classes with a directional force in equilibrium. We attribute this deviation from the equilibrium prediction (see Materials and Methods) to demography, in particular to recent population growth (Johri et al., 2020). Furthermore, the ratios do not differ significantly between GC-conservative mutations (a, f), yet among the GC-changing mutations they are slightly higher for transversions (d, e) than for transitions (b, c) (figure S1). This might be due to a difference in the strength of gBGC between transitions and transversions, which has been shown to be the case in other organisms (Lartillot, 2012; Bergman & Schierup, 2021).

In summary, our analysis of nucleotide polymorphism and divergence patterns along chromosomes show that linked selection can only account for part of the pattern. Varying divergence over polymorphism ratios in different mutation classes indicate an additional force, likely the non-adaptive directional force of gBGC.

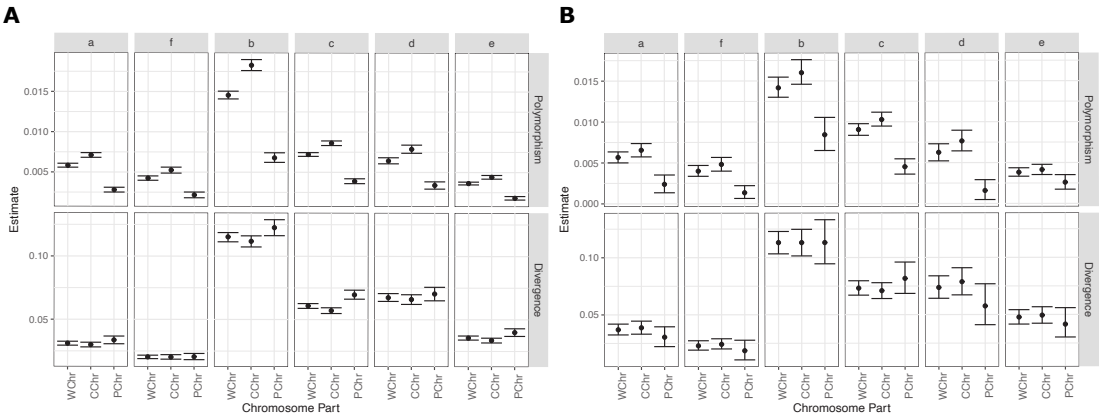

Figure 2: Scaled mutation rate (polymorphism) and divergence estimates of 5SI for the six mutation classes from the different parts of A) autosomes and B) X chromosome, where a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions. WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

Table 2: Divergence over polymorphism ratios for GC-conservative (a and f) and GC-changing (b, c, d, e) mutations for different parts of the autosomes and X-chromosome (95% CIs in brackets). WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

	Autosome		X	
	GC-conservative	GC-changing	GC-conservative	GC-changing
WChr	5.104 (4.846, 5.394)	8.774 (8.452, 9.095)	6.191 (5.280, 7.263)	9.246 (8.410, 10.195)
CChr	4.062 (3.819, 4.334)	6.854 (6.565, 7.130)	5.534 (4.715, 6.502)	8.201 (7.353, 9.184)
PChr	10.908 (9.659, 12.288)	19.065 (17.570, 20.414)	13.142 (8.606, 21.736)	17.167 (13.276, 21.695)

Diversity estimates between autosomes and the X chromosome

Under neutral equilibrium conditions (*e.g.,* no mutational or effective population size difference between sexes) the expected divergence ratio of X over autosomes should be one and the expected polymorphism ratio $3/4 = 0.75$. However, previously reported estimates of the X/A neutral site diversity in the ancestral populations of *Drosophila melanogaster* are approximately one (Jackson et al., 2017). This was explained by linked selection, specifically by background selection (BGS): a higher recombination rate in the X chromosome should counteract the effect of BGS leading to less reduction in diversity. To support this argument, it was reported that the observed ratio of X/A diversities are recovered when BGS is modelled with the estimates of the distribution of fitness effect of deleterious mutations (Charlesworth, 2012; Comeron, 2014). Furthermore, when regions with similar effective recombination rates are compared, the ratio of mean X/A diversity values was shown to be close to the expected value of 0.75 (Vicoso & Charlesworth, 2009; Campos et al., 2013, 2014). However, our analyses above showed that mutation classes might be affected by different forces, thus we investigated the X/A polymorphism and the divergence ratios separately for each class.

Except for the peripheral regions, both polymorphism and divergence ratios are generally higher than their expected values (0.75 and 1, respectively, table 3). Yet, the deviation in X/A polymorphism ratios can not be only driven by BGS as suggested before, both because there are differences between mutation classes and the mutation rates are also slightly higher for the X chromosome compared to autosomes. Can these deviations be explained by mutation rate differences between X and autosomes? In the case of a pure mutation rate effect, polymorphism and divergence estimates of the chromosomes increase or decrease proportionally, so that the ratio between these two estimates should be unaffected. Thus without the effect of any directional force, we would expect $X_{pol}/A_{pol} = 0.75\, X_{div}/A_{div}$ (See also equations 5 and 6).

This expectation holds for GC-conservative mutations (a, f) (figures 3, S2): a and f mutations seem to evolve under purely neutral forces and the higher X/A polymorphism ratio can be explained by a high X-chromosomal mutation rate in these nucleotide classes without invoking the effect of linked selection. Compared to autosomes, the X chromosome has relatively higher AT-to-GC mutation rates (c, e). This finding is unsurprising, as we have previously reported a reduced mutation bias towards AT on the X chromosome. However, nucleotide classes without mutation bias (a, f) also exhibit slightly elevated mutation rates in the X chromosome. This suggests that factors other than mutational bias contribute to the differences in mutation rates between chromosomes.

While the GC-conservative mutation classes follow the neutral expectation $X_{pol}/A_{pol} = 0.75\, X_{div}/A_{div}$, the GC-changing mutation classes differ (figures 3, S2). The different pattern for the GC-changing mutations suggests that gBGC might be affecting the molecular evolution patterns

of X and autosomes differently. Considering the GC-changing mutations separately, the differences between X and autosomes seem mostly caused by gBGC acting on transitions (b and c) and also on telomeric and centromeric regions.

In summary, mutation rates differ between autosomes and the X chromosome. Upon accounting for these differences, GC-conservative mutations conform to neutral expectations. But the pattern of GC-changing mutations suggests gBGC in shaping chromosomal disparities. Thus, when comparing the evolution of autosomes and the X chromosome, analyses should either be restricted to GC-conservative mutation classes or gBGC needs to be accounted for.

Table 3: X/A ratios for polymorphism and divergence estimates of 5SI for each mutation class from the different parts of chromosomes (95% CIs in brackets); a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions; WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

	Polymorphism			Divergence		
	WChr	CChr	PChr	WChr	CChr	PChr
a	0.97 (0.85, 1.10)	0.92 (0.79, 1.06)	0.85 (0.48, 1.28)	1.19 (1.03, 1.36)	1.29 (1.09, 1.51)	0.90 (0.66, 1.21)
f	0.94 (0.79, 1.11)	0.92 (0.76, 1.10)	0.62 (0.28, 1.04)	1.12 (0.93, 1.36)	1.19 (0.97, 1.46)	0.90 (0.51, 1.37)
b	0.97 (0.88, 1.07)	0.87 (0.79, 0.96)	1.24 (0.96, 1.57)	0.98 (0.89, 1.08)	1.01 (0.91, 1.13)	0.92 (0.76, 1.10)
c	1.26 (1.15, 1.37)	1.19 (1.09, 1.31)	1.16 (0.92, 1.43)	1.20 (1.10, 1.32)	1.24 (1.13, 1.38)	1.17 (0.98, 1.40)
d	0.98 (0.80, 1.15)	0.97 (0.81, 1.14)	0.47 (0.15, 0.88)	1.09 (0.94, 1.25)	1.19 (1.02, 1.39)	0.82 (0.58, 1.11)
e	1.05 (0.91, 1.21)	0.94 (0.79, 1.10)	1.44 (0.95, 1.96)	1.34 (1.16, 1.53)	1.47 (1.26, 1.71)	1.04 (0.75, 1.42)

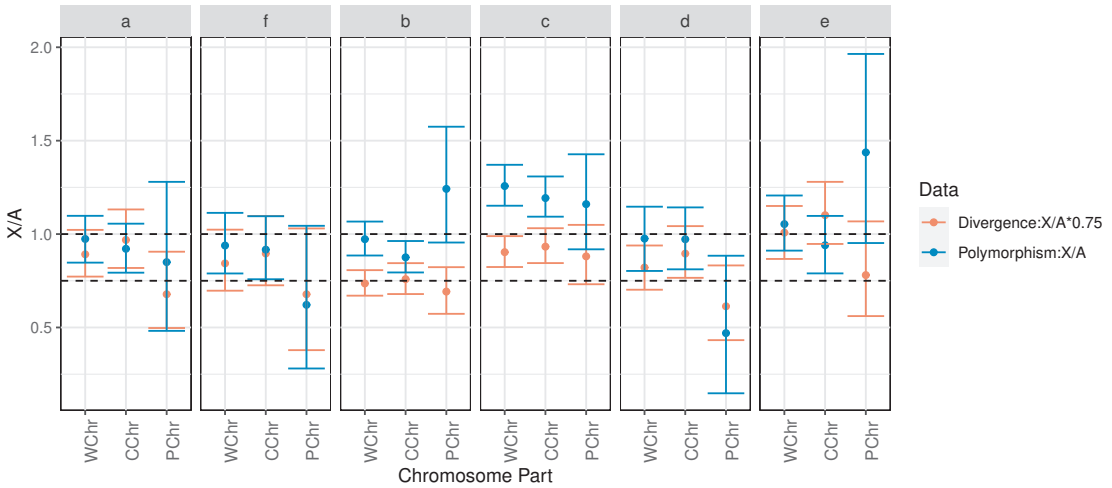

Figure 3: X/A ratios for polymorphism and divergence estimates. The divergence ratios are multiplied with 0.75 to account for the expectation of $X_{pol}/A_{pol} = 0.75\ X_{div}/A_{div}$ among mutation classes; a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions; horizontal dashed lines correspond to the values of 0.75 and 1. WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

Diversity estimates in relation to recombination rates

Given the lack of variation in divergence levels, the variation in polymorphism levels along the genome has been explained by the effect of linked selection, thus with variation in N_e in *Drosophila*

species (Begun & Aquadro, 1992). The effect of linked selection should not differ among mutation classes. Contrary to this expectation, we observed distinct patterns among mutation classes and along chromosomes that we attributed to gBGC (table 2). As both gBGC and linked selection should be tied to recombination, we next investigated the diversity estimates of six mutation classes for different recombination rates. For this, we combined introns with similar recombination rates and compared results among them. We created four bins with approximately equal numbers of observations and also performed the binning after excluding telomeres and centromeres. The mean recombination rates within bins are approximately equal for X and autosomes (table S1).

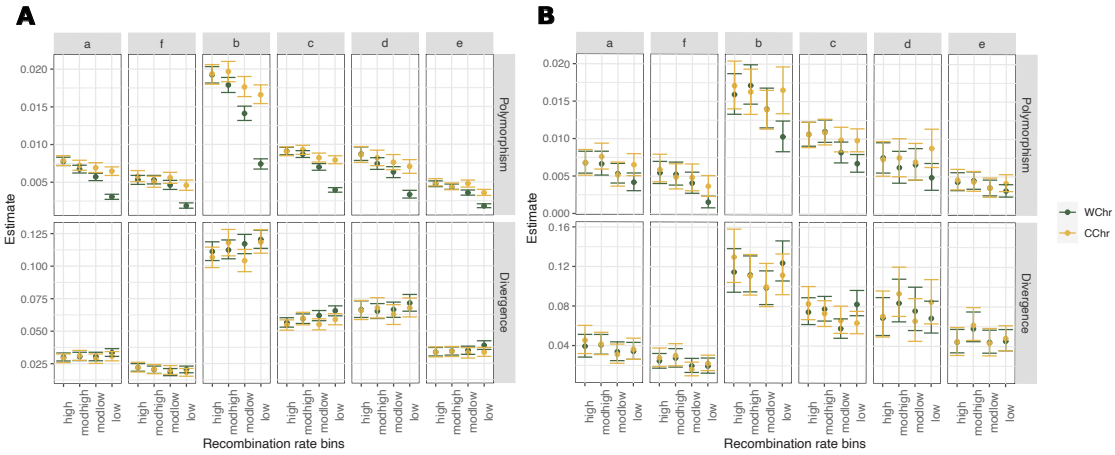

Figure 4: Scaled mutation rate (polymorphism) and divergence estimates of 5SI for the six mutation classes from the different recombination rates of A) autosomes and B) X chromosome, where a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions. Introns are binned by recombination rate before (green, WChr: Whole chromosome) and after excluding telomeres and centromeres (yellow, CChr: Central regions of the chromosome arms).

The relative relationship between the estimates of mutation classes follows the same patterns reported before: higher rates for transitions and GC-to-AT mutations (figure 4). Among the recombination bins including telomeres and centromeres, polymorphism estimates decrease with decreasing recombination rate, which is much more pronounced for autosomes. However, the variation in divergence between recombination bins is not significant, showing once again that mutation rate variation does not significantly contribute to the positive correlation between polymorphism and recombination in *Drosophila*. When peripheral regions of the chromosome arms with very low recombination rates are excluded, the decrease in polymorphism levels diminishes relatively. This suggests that the forces contributing to the association between polymorphism and recombination are mainly caused by differences between the central and peripheral regions of the chromosome arms.

Next we asked to what extent the polymorphism-recombination relationship is caused by linked selection and possibly by gBGC. We factored out mutational differences by calculating the ratio of divergence over polymorphism. In whole chromosomes we find that the lowest recombination rate class has significantly higher ratios than all other classes (figure 5). Comparing GC-conservative (a, f) and changing (b, c, d, e) mutations, we note that this effect is strongest in GC-changing mutations. Thus gBGC seems to contribute to the positive correlation between nucleotide polymorphism

and local rates of recombination. The other recombination classes differ little from each other. After excluding telomeres and centromeres, variation is only significant for the lowest recombination bin of GC-changing mutations and surprisingly, the difference between GC-changing transitions and transversions is also lost. In contrast, no differences among recombination classes can be found for GC-conservative mutations.

In summary, divergence over polymorphism ratios differ among recombination classes only when the peripheral regions of chromosome arms are included. Variation among recombination classes within the central regions is low. Particularly, there is no trend toward increasing ratios (and thus presumably to decreasing effective population sizes due to linked selection) across different recombination rate classes within the central chromosome arms in GC-conservative mutation classes, while such a trend is discernible but barely significant in GC-changing mutation classes.

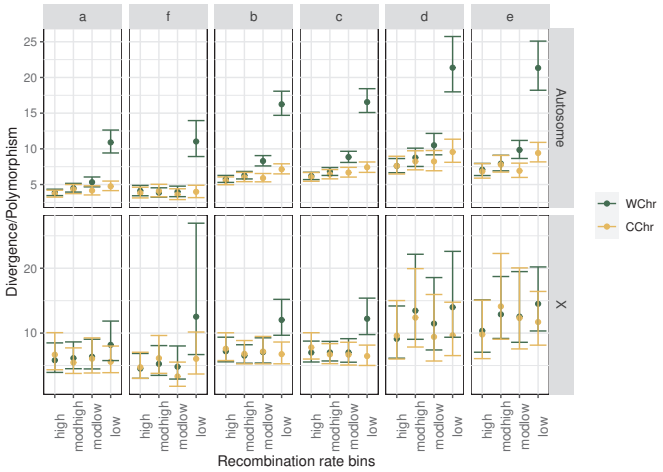

Figure 5: Divergence over polymorphism ratios for all six mutations classes from autosomes and X-chromosome, where a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions. Introns are binned by recombination rate before (green, WChr: Whole chromosome) and after excluding telomeres and centromeres (yellow, CChr: Central regions of the chromosome arms).

Lastly, we checked if the X/A diversity ratios among recombination classes conform to the expectation of $X_{pol}/A_{pol} = 0.75\, X_{div}/A_{div}$. As above, polymorphism ratios are generally higher than or equal to 3/4 and the patterns change between GC-changing and conservative classes (figure 6). For GC-conservative mutations, the observed ratio is totally explained by the mutation rate differences when telomeres and centromeres are excluded and deviates only slightly for the lowest recombination bin when they are included. It seems that the effect of background selection is creating a difference in X and autosome variation patterns only in centromeres/telomeres. For GC-changing mutations, there are again deviations from the neutral expectation, stronger towards low recombination rates and for transitions (figure S3). When telomeres and centromeres are excluded, the deviation is still observed for the lowest recombination bin (figure 6) due to transitions (figure S4). These results are in line with the findings above: the effect of gBGC on X and autosome differs mostly for transitions and in very low recombining regions.

Previous studies showed that the X/A diversity ratio is close to the neutral expectation when regions with equal effective recombination rates are compared (*i.e.,* the rates for the X chromosome

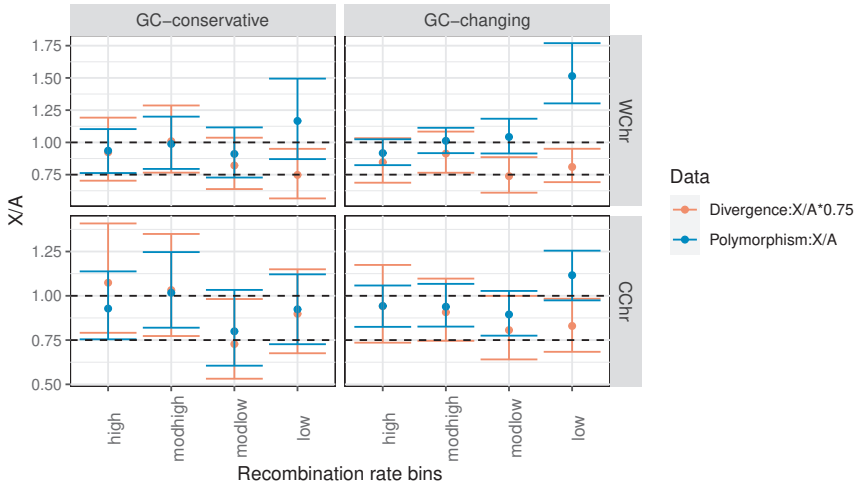

Figure 6: X/A ratios for polymorphism (blue) and divergence (red) estimates of GC-conservative (a and f) and changing (b, c, d, e) mutations. Introns are binned by recombination rate before excluding telomeres and centromeres (top panel, WChr: Whole chromosome) or after excluding telomeres and centromeres (lower panel, CChr: Central regions of the chromosome arms). The divergence ratios are multiplied with 0.75 to account for the expectation of $X_{pol}/A_{pol} = 0.75\ X_{div}/A_{div}$ among mutation classes. Horizontal dashed lines correspond to the values of 0.75 and 1.

are 4/3 of the rates for autosomes due to lack of recombination in male *Drosophila*) with an exception of regions with very low recombination (Vicoso & Charlesworth, 2009; Campos et al., 2013, 2014). They suggested that this further supports BGS as the main driver of the high X/A diversity ratio when whole chromosomes are considered and as long as regions with similar recombination rates are compared, expectations for neutral diversity should be met. However, we observed that patterns differ between GC-conservative and GC-changing mutations. Specifically, after accounting for mutational differences, there is no deviation from neutral equilibrium expectations in the GC-conservative mutations in the central regions of chromosome arms. Therefore BGS appears not to be the driver of chromosomal differences, except at telomeres and centromeres, while the effect of gBGC should be considered.

Strength of gBGC

As we find evidence for the effect of gBGC in diversity patterns, we next inferred its strength, quantified as $B = 4N_e b$, using the ML estimator of Vogl & Bergman (2015). We obtained estimates either jointly for all GC-changing mutations (B_{GC}) or separately for GC-changing transitions (B_{Ts}) and transversions (B_{Tv}) from SFS constructed based on segregating GC-frequency. We tested whether separately considering transitions and transversions improved the fit by a likelihood ratio test, where we compared the likelihood of B_{GC} to the sum of the likelihoods of B_{Tv} and B_{Ts}. The difference between the estimates of autosomes and the X chromosome or between the peripheral and central parts of chromosomes are evaluated in a similar manner.

The force favoring GC is weak, *i.e.,* about $B \approx 0.5$, but differs significantly from zero both in autosomes and the X chromosome (table 4). Estimates are also significantly different from zero if transitions and transversions are analyzed separately (table 5). For central autosomal arms, B_{Tv} is

significantly stronger than B_{Ts} (LRT $\chi^2_{df=1} = 8.916, p = 0.0028$), while for peripheral regions, the difference is not significant (LRT $\chi^2_{df=1} = 0.296, p = 0.586$). B_{Tv} estimates do not differ between the central and peripheral parts of the chromosome arms (LRT $\chi^2_{df=1} = 0.006, p = 0.938$), while B_{Ts} estimates do (LRT $\chi^2_{df=1} = 8.096, p = 0.004$). Conversely, for the X chromosome, B_{Ts} and B_{Tv} differ significantly at the peripheral region of the chromosome arms, with a higher B_{Tv} (LRT $\chi^2_{df=1} = 5.785, p = 0.016$), while at the central part the value for all GC-changing mutations fits the data better (LRT $\chi^2_{df=1} = 1.346, p = 0.246$). This is explained by an increasing B_{Tv} value towards telomeres and centromeres, as B_{Ts} does not change significantly along the X chromosome (LRT $\chi^2_{df=1} = 0.121, p = 0.728$), while B_{Tv} does (LRT $\chi^2_{df=1} = 8.414, p = 0.003$).

On both autosomes and the X chromosome, there is an increase in the overall strength of gBGC (B_{GC}) towards the telomeres and centromeres (table 4). This is due to a significant increase in B_{Ts} for autosomes, while B_{Tv} shows a significant increase for the X chromosome. To assess whether these changes along the chromosomes cause significant differences between the X chromosomal and autosomal estimates of gBGC, we compared pooled data to their estimates via LRT. For the central part of the chromosome arms, the difference between the X chromosome and autosomes is significant due to the differences in B_{Ts} (LRT $\chi^2_{df=1} = 4.228, p = 0.039$), while for the peripheral regions, differences in B_{Tv} cause a significant deviation (LRT $\chi^2_{df=1} = 6.529, p = 0.010$). These results are consistent with those of diversity patterns, both in comparisons along chromosomes and between chromosomes.

We also estimated B values while accounting for the effect of demography through addition of correction parameters (r_y) that are obtained from the neutral SFS, *i.e.,* SFS of GC-conservative mutations (see Materials and Methods). We note that this decreases the statistical power. Nonetheless, all estimates remain significantly different from zero, and the values fall within a similar range as the estimates without correction (see Figure S5 and Tables S4 and S5). Furthermore, we still observe an increase of B towards telomeres and centromeres in B_{Ts} for autosomes and in B_{Tv} for X chromosomes, however, these trends are no longer statistically significant, likely due to the reduced power (figure S5).

Table 4: B values inferred from the SFS of GC-changing mutations (B_{GC}), for autosomes, the X chromosome, and autosomes and X pooled. Estimates are either given separately for central and peripheral regions of the chromosome arms (CChr and PChr, respectively) or for whole chromosomes (WChr). All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	WChr	CChr	PChr
Autosomes	0.504	0.477	0.650
X	0.597	0.564	0.853
Pool	**0.515**	**0.488**	**0.668**

That our estimates of B do not decrease near telomeres and centromeres compared to the central part of the chromosome arms, even after accounting for demographic disequilibrium, requires an explanation. $B = 4N_e b$ measures the strength of the directional force scaled by the effective

Table 5: B values inferred from the SFS of GC-changing transitions (B_{Ts}) and transversions (B_{Tv}), for autosomes, the X chromosome, and autosomes and the X pooled. Estimates are either given separately for central and peripheral regions of the chromosome arms (CChr and PChr, respectively) or for whole chromosomes (WChr). All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	WChr		CChr		PChr	
	Transitions	Transversions	Transitions	Transversions	Transitions	Transversions
Autosome	0.457	0.605	0.417	0.606	0.677	0.596
X	0.613	0.559	0.625	0.425	0.529	1.781
Pool	0.476	**0.600**	0.443	**0.585**	**0.662**	0.682

population size. The conversion bias b depends on the average length of the conversion tract, the repair bias towards GC, and the recombination rate per site per generation. Notably, our recombination rate estimates rely on only crossover (CO) events. As CO rates and the effective population size decreases towards telomeres and centromeres (see the mutation classes a and f in the autosomes in figure 5), the conversion bias b needs to increase disproportionately to explain the patterns in B values. This suggests that b might be governed by a more intricate interplay of factors than a simple relationship with CO rates.

To show the relationship between gBGC and the recombination (CO) rate, we calculated B from the SFS data binned according to CO rates. Although there is a slight increase in B towards low recombining regions, when telomeres and centromeres are included, the estimates do not differ significantly among recombination bins (figure S6). Importantly, estimates inferred without recombination binning (table 4) provide a better fit to the data (LRT $\chi^2_{df=3} = 2.786, p = 0.426$ for autosomes; LRT $\chi^2_{df=3} = 3.071, p = 0.380$ for the X chromosome). Furthermore, the B estimates obtained without assuming demographic equilibrium also do not exhibit any positive or linear relationship with CO rates. There is still a slight, non-significant increase in B towards low-recombining regions (figure S7). When telomeres and centromeres are included, estimates from the lowest recombination rate classes are similar to those from the highest recombination rates.

Our inference of gBGC strength so far relies on polymorphic spectra and is thus limited to currently segregating alleles. On the other hand, the general GC content including monomorphic sites is influenced also by earlier events that already fixed. Using the GC content thus promises more power for inference. A slight but significant negative correlation between GC content and CO rates is observed for the 5SI regions when analyzing whole chromosomes (Spearman's $\rho = -0.048, p < 0.001$ for autosomes; Spearman's $\rho = -0.140, p < 0.001$ for the X chromosome). After excluding telomeres and centromeres, the significance is lost for autosomes and reduced for the X chromosome (Spearman's $\rho = -0.018, p = 0.192$ for autosomes; Spearman's $\rho = -0.116, p = 0.002$ for the X chromosome).

While the CO rate and the strength of gBGC are expected to be indirectly related, the relationship between GC content and gBGC is expected to be direct. Consistent with our observation that gBGC is relatively constant along the chromosomes, the GC content of the *Drosophila* genome also shows little variation. Yet, when examining the differences in base composition, we may still expect a weak correlation between B values and GC content. In order to investigate this, it is necessary to group introns into bins according to their GC content. However, estimating the strength of gBGC from data binned according to its own GC content might create a bias. To avoid this dependence, we created five bins with approximately equal sizes by using the mean GC content of fourfold degenerate sites from the same gene where the short introns are located, as the GC

content of four-fold degenerate sites (FFDS) and 5SI is significantly and positively correlated across all chromosomes (Spearman's $\rho = 0.257, p < 0.001$, Spearman's $\rho = 0.273, p < 0.001$ for autosomes and X chromosome, respectively). Most of the data clustered in intermediate levels of GC content (figure S8), thus the GC content ranges of the bins were not equal (table S2). Furthermore the mean CO rates were similar among GC-bins ($\approx$ 2.27 cM/Mb and $\approx$ 2.56 cM/Mb for autosomes and X chromosome, respectively). Even though a relationship between GC content, thus gBGC, and CO rates is expected, we once more fail to observe it.

The estimates of B_{Ts} and B_{Tv} do not significantly differ from each other, *i.e.,* B_{GC} fits data better, for both chromosomes and for all GC-bins. More crucially, estimates from GC-bins explain our data better than the whole chromosome estimates (LRT $\chi^2_{df=4} = 30.325, p < 0.001$ for autosomes; LRT $\chi^2_{df=4} = 21.146, p < 0.001$ for X chromosome). Among the GC-bins, B_{GC} is higher for highest GC content, yet the difference among the other four classes is not significant (figure 7A). Due to the relatively uniform base composition along chromosomes, we expected little power to detect co-variation in B values, nevertheless we detected such an association. Additionally, while the strength of the directional force, *i.e.,* gBGC, increases with GC content, the mutation bias towards AT is independent of it (figure 7B). Accounting for the demographic disequilibrium had no effect on the patterns between GC-bins, with the only impact being on the absolute values of B (figure S9).

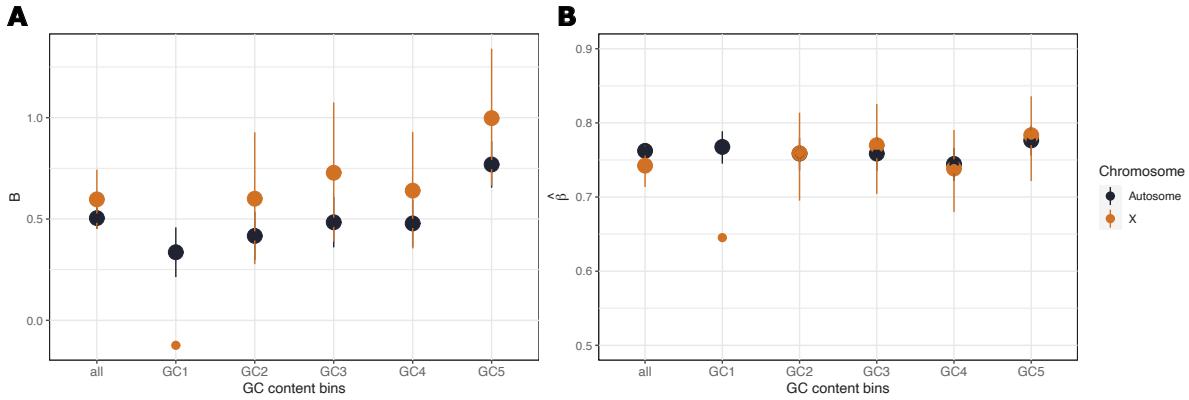

Figure 7: A) B_{GC} values inferred from the SFS of GC-changing mutations for autosomes (black) and X chromosome (brown). The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test. B) Mutation bias estimated conditional on B for for autosomes (black) and X chromosome (brown). Estimates are given for all introns and for introns binned by the mean GC content of the FFDS of the same genes. GC content increases from GC1 to GC5 and the ranges are given in S2

Comparison to Drosophila simulans

So far we used polymorphism data from *D. melanogaster* and divergence data between *D. melanogaster* and *D. simulans*. As the shared polymorphism between the two species is negligible, we also analyzed the polymorphic spectrum of *D. simulans* to compare patterns of the different mutation classes between the species.

As in *D. melanogaster*, the inferred strength of gBGC is about $B = 0.5$ and significantly different from $B = 0$ (table 6). Estimates are higher for GC-changing transversions than for GC-changing transitions (LRT $\chi^2_{df=1} = 64.078, p < 0.001$, LRT $\chi^2_{df=1} = 4.548, p = 0.0329$ for autosomes

and X, respectively). Importantly, divergence over polymorphism ratios are higher for GC-changing mutations than for GC-conservative ones and increase with the strength of gBGC, again as in *D. melanogaster* (figure S10).

Furthermore, the GC content of four-fold degenerate and 5SI sites of the same gene are also correlated significantly and positively, as in *D. melanogaster* (Spearman's $\rho = 0.267, p < 0.001$, Spearman's $\rho = 0.283, p < 0.001$ for autosomes and the X chromosome, respectively). Thus, we also grouped the SFS into five equally sized bins depending on the background FFDS base composition and inferred B (table S3). The estimates are significantly greater than $B = 0$ for all GC-bins in autosomes and for the two bins with the highest GC content in the X chromosome (figure 8A). Among the significant estimates, the B values increase with GC content, while the mutation bias towards AT remains constant (figure 8B). Overall, patterns in *D. simulans* are similar to those in *D. melanogaster*. Once again, B values obtained after accounting for the disequilibrium follow the same patterns (table S6). However, the difference between B_{Tv} and B_{Ts} estimates and the estimates from the low GC bins on the X chromosome are not significant anymore (figure S11, S12).

In summary, gBGC is present and affecting neutral sequence variation similarly in *Drosophila melanogaster* and *D. simulans*. This manifests in varying GC content along chromosomes, patterns of skewed site frequency spectra, and deviations in divergence to polymorphism ratios of GC-changing mutations compared to GC-neutral mutations.

Table 6: B values inferred from the SFS of all GC-changing mutations (B_{GC}), GC-changing transitions (B_{Ts}) and transversions (B_{Tv}) for autosomes, the X chromosome, and pooled data (autosomes+X) of *D. simulans* population. All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	B_{GC}	B_{Tv}	B_{Ts}
Autosomes	0.459	0.763	0.311
X	0.345	0.588	0.232
Pool	**0.447**	**0.745**	**0.303**

Discussion

Identifying the different nonadaptive evolutionary forces that act on putatively neutral sequences and describing their effects provides a better understanding of genome evolution patterns. These nonadaptive evolutionary forces can then be incorporated into null-models to quantify their influence in the genome. By comparing the neutral patterns to genomic regions of functional importance, such models allow for the detection of adaptive forces. In such null-models, the effect of linked selection has been included in addition to mutation and demography (Zeng & Charlesworth, 2010; Comeron, 2017; Johri et al., 2020), as it has been shown to shape patterns of genome variation in many organisms. More recently, the effect of GC-biased gene conversion (gBGC) has been demonstrated in many taxa (Pessia et al., 2012; Glémin et al., 2014; Galtier, 2021), including *Drosophila* (Jackson & Charlesworth, 2021). Failing to account for gBGC, when it is present, has been shown

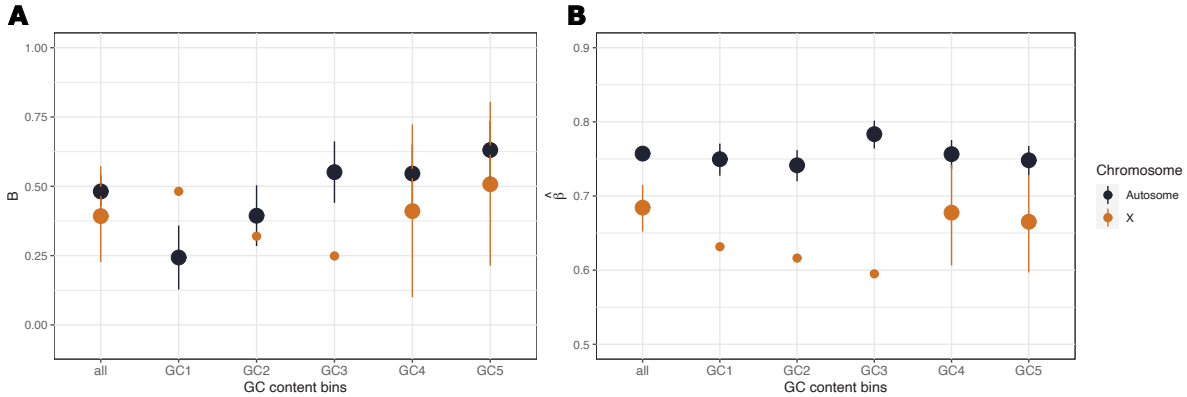

Figure 8: Results from GC content binned data of *D. simulans* population. A) B_{GC} values inferred from the SFS of GC-changing mutations for autosomes (black) and X chromosome (brown). The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test. B) Mutation bias estimated conditional on B for for autosomes (black) and X chromosome (brown). Estimates are given for all introns and for introns binned by the mean GC content of the FFDS of the same genes. GC content increases from GC1 to GC5 and the ranges are given in S3

to lead to biased inference of selective and demographic forces (Bolívar et al., 2018; Pouyet et al., 2018), or lead to false interpretations about the effect of nonadaptive forces in shaping neutral sequence patterns (Bolívar et al., 2016; Boman et al., 2021). Thus, it is important to incorporate gBGC into the model when inferring adaptive forces and to use GC-conservative mutations for a more accurate representation of the effects of other nonadaptive factors governing neutral genome sequence evolution.

Using presumably neutrally evolving sequences in *Drosophila*, we find a pervasive influence of gBGC on patterns of neutral sequence variation in both *Drosophila melanogaster* and *D. simulans* that shows in variable GC content along chromosomes, correlated skewed polymorphism patterns, deviation of divergence to polymorphism ratios from predictions assuming only mutation and drift, and differences among transition and transversion mutations and between autosomes and the X chromosome. On the other hand, patterns in GC-conservative mutations show that most predictions of the neutral theory are borne out, while many of the results formerly attributed to linked selection seem to actually be caused by gBGC.

Analysis of GC-changing mutation classes shows the presence of a directional force attributable to gBGC of about $B \approx 0.5$. It operates in all chromosomal regions of autosomes and the X chromosome and in GC-changing transitions and transversions, albeit with slightly varying strength among regions. Between autosomes and the X chromosome the pattern is complicated: transitions in the central regions and transversions in the peripheral regions are higher in the X chromosome (table 5), yet after accounting for the lack of equilibrium, differences are not significant (figure S5). Along chromosomes, the strength of gBGC slightly increases towards telomeres and centromeres, where crossover (CO) rates are low (Comeron et al., 2012). This relationship between CO rates and the strength of gBGC is also reflected in the negative relationship between 5SI GC content and CO rates.

In *Drosophila*, noncrossover (NCO) and crossover (CO) rates are negatively correlated (Langley et al., 2000; Comeron et al., 2012) and NCO rates exhibit a more uniform distribution along the

chromosomes compared to CO rates (Comeron et al., 2012; Miller et al., 2016). Our data indicate a higher or similar values of directional force $B = 4N_e b$ towards peripheral regions of chromosome arms where CO rates and effective population sizes N_e are low (N_e is here estimated independently from the GC-conservative mutations). This combination of observations suggests that the varying strength of the conversion bias b is not only associated with COs but also with NCOs in *Drosophila*. Such an association would explain both the negative relationship between 5SI GC content and CO rate and the relatively uniform strength of gBGC (and as a corollary the relatively uniform GC content in 5SI), except between the central and peripheral part of the chromosome arms. Previous studies failed to find a negative association between the GC content of introns and the CO rate in *D. melanogaster*, but rather reported weak positive correlations when considering whole genome GC content (Marais et al., 2003; Singh et al., 2005a). This might be due to an incomplete annotation of the reference genome, due to data excluding telomeres and centromeres, or because sites affected by directional selection were included. To our knowledge, only two studies reported a negative correlation between non-coding GC content and recombination in the X chromosome of *D. melanogaster* (Singh et al., 2005b; Campos et al., 2013). They also suggested higher gBGC in regions of low recombination as a possible explanation among others, however, discarded this possibility due to the apparent absence of such a relationship on the autosomes. As we observe such a negative relationship between crossover rate and GC content on both autosomes and the X chromosome, however, we uphold this hypothesis.

Some of the patterns of diversity and skew in the site frequency spectra observed in our study could also be explained by a recent change in the mutation bias instead of gBGC. We next summarize the arguments for gBGC using our data and analyses: while the genome of *D. melanogaster* has become more AT-rich compared to the ancestral state (Kern & Begun, 2005; Jackson & Charlesworth, 2021), which may explain a GC-skewed polymorphic SFS via a change in mutation bias, that of *D. simulans* has not (Jackson & Charlesworth, 2021), but nevertheless shows similarly skewed polymorphic SFS and diversity patterns as *D. melanogaster* (table 6, figures 8, S10). Furthermore, both in *D. melanogaster* and in *D. simulans* the GC proportion varies over the genome. The GC content of introns is correlated along chromosomes with that of fourfold degenerate sites (FFDS), pointing to a common mechanism. Using the polymorphic site frequency spectra (SFS), we inferred gBGC of varying strength in the 5SI correlated with GC content (figures 7, 8, S9, S12), but a rather uniform mutation bias that cannot explain variation in GC proportion. In addition, unlike Jackson & Charlesworth (2021), who partitioned their data based on the GC content of 5SI, which may introduce bias, we instead utilized the GC content of FFDS. Our observation of stronger B values for higher GC content provides further support for the common mechanism being gBGC. Thus, our evidence strongly points to gBGC as the more plausible explanation over a change in mutation bias.

Although the directional force of gBGC is within the nearly neutral range (Tachida, 1991), it has an impact on diversity patterns and can lead to false interpretations of genome evolution if not properly accounted for. GC-conservative mutation classes are not affected by gBGC and therefore suited to infer the effects of linked selection. We infer lower effective population sizes towards telomeres and centromeres than in central regions. This correlates with the overall pattern of CO rates. On the other other hand, variation in the local effective population size N_e in the central chromosomal regions is small and not significantly correlated with CO rates (figure 5). This has important consequences for the current efforts building null models: Within the central regions of chromosomes the CO map has little predictive value explaining the observed sequence patterns.

Comparing autosomes and the X chromosomes in *D. melanogaster* using GC-conservative

mutations shows that, while the overall mutation rate is higher on the X, the effective population size of the X is about 3/4th that of the autosomes in the central region of chromosome arms (table 3 and figures 3,6, S2). The exact biological mechanism for the higher X mutation rate is unclear. Higher female mutation rate or differences in heterogametic male X chromosome, like dosage compensation (Gupta et al., 2006; Lucchesi & Kuroda, 2015) and distinct repair properties, might cause increase in mutation rates. In any case, our finding is important for studies comparing evolutionary patterns between the X chromosome and autosomes: obviously the ratio of the effective population size on the X chromosome to that of the autosomes is just as expected from the neutral theory, given the differences in effective population sizes of autosomes and the X chromosome. Failing to differentiate among mutation classes might obscure this simple pattern.

Differentiating between mutation classes allows us to tease apart the influence of various population genetic forces relevant for neutral sequence evolution patterns in *Drosophila*: the influence of mutation rates in different mutation classes and their variation, the influence of differences in recombination rates and linked selection, and the influence of gBGC. None of these individual forces dominates, rather they act jointly and interdependently. In our study, we took the joint influence of all weak forces on neutral sequence evolution in *Drosophila* into account. We believe that doing so will also improve the study of neutral and nearly neutral sequence evolution in other species.

References

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*, *136*, 927–35.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics*, *139*, 1067–1076.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, *437*, 1149–1152. doi:10.1038/nature04107.

Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, *356*, 519–520. doi:10.1038/356519a0.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLOS Biology*, *5*, 1–26. doi:10.1371/journal.pbio.0050310.

Bergman, J., Betancourt, A. J., & Vogl, C. (2017). Transcription-associated compositional skews in *Drosophila* genes. *Genome Biology and Evolution*, *10*, 269–275. doi:10.1093/gbe/evx200.

Bergman, J., & Schierup, M. H. (2021). Population dynamics of GC-changing mutations in humans and great apes. *Genetics*, *218*, iyab083. doi:10.1093/genetics/iyab083.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., & Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, *20*. doi:10.1186/s13059-018-1613-z.

Bolívar, P., Mugal, C., Nater, A., & Ellegren, H. (2016). Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill-Robertson interference, in an avian system. *Molecular Biology and Evolution*, *33*, 216–227. doi:10.1093/molbev/msv214.

Bolívar, P., Mugal, C. F., Rossi, M., Nater, A., Wang, M., Dutoit, L., & Ellegren, H. (2018). Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. *Molecular Biology and Evolution*, *35*, 2475–2486. doi:10.1093/molbev/msy149.

Boman, J., Mugal, C. F., & Backström, N. (2021). The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes. *Genome Biology and Evolution*, *13*. doi:10.1093/gbe/evab064.

Borges, R., Szöllősi, G. J., & Kosiol, C. (2019). Quantifying GC-biased gene conversion in great ape genomes using Polymorphism-Aware models. *Genetics*, *212*, 1321–1336. doi:10.1534/genetics.119.302074.

Campos, J. L., Halligan, D. L., Haddrill, P. R., & Charlesworth, B. (2014). The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *31*, 1010–1028. doi:10.1093/molbev/msu056.

Campos, J. L., Zeng, K., Parker, D. J., Charlesworth, B., & Haddrill, P. R. (2013). Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *30*, 811–823. doi:10.1093/molbev/mss222.

Chamary, J., Parmley, J., & Hurst, L. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Review Genetics*, *7*, 98–108. doi:10.1038/nrg1770.

Charlesworth, B. (2012). The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the drosophila X chromosome. *Genetics*, *191*, 233–46. doi:10.1534/genetics.111.138073.

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*, 1289–303.

Clemente, F., & Vogl, C. (2012). Unconstrained evolution in short introns? - an analysis of genome-wide polymorphism and divergence data from *Drosophila*. *Journal of Evolutionary Biology*, *25*, 1975–1990. doi:10.1111/j.1420-9101.2012.02580.x.

Comeron, J. M. (2014). Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genetics*, *10*, e1004434. doi:10.1371/journal.pgen.1004434.

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. *Philosophical Transactions of the Royal Society B*, *372*. doi:10.1098/rstb.2016.0471.

Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, *8*, 1–21. doi:10.1371/journal.pgen.1002905.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*, 3133–3157. doi:10.1111/mec.12796.

Dohet, C., Wagner, R., & Radman, M. (1985). Repair of defined single base-pair mismatches in *Escherichia coli*. *Proceedings of the National Academy of Sciences, USA*, *82*, 503–505. doi:10.1073/pnas.82.2.503.

Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, *10*, 285–311. doi:10.1146/annurev-genom-082908-150001.

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*, 693–709. doi:10.1093/genetics/133.3.693.

Galtier, N. (2021). Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Community Journal*, *1*. doi:10.24072/pci.genomics.100012.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., & Duret, L. (2018). Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Molecular Biology and Evolution*, *35*, 1092–1103. doi:10.1093/molbev/msy015.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in north american *Drosophila melanogaster* show signatures of soft sweeps. *Plos Genetics*, *11*, e1005004. doi:10.1371/journal.pgen.1005004.

Glémin, S., Clément, Y., David, J., & Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics*, *30*, 263–270. doi:10.1016/j.tig.2014.05.002.

Gupta, V., Parisi, M., Sturgill, D., Nuttall, R., Doctolero, M., Dudko, O. K., Malley, J. D., Eastman, P. S., & Oliver, B. (2006). Global analysis of X-chromosome dosage compensation. *Journal of Biology*, *5*. doi:10.1186/jbiol30.

Haddrill, P. R., & Charlesworth, B. (2008). Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biology Letters*, *4*, 438–441. doi:10.1098/rsbl.2008.0174.

Haddrill, P. R., Charlesworth, B., Halligan, D. L., & Andolfatto, P. (2005). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, *6*, R67. doi:10.1186/gb-2005-6-8-r67.

Halligan, D. L., & Keightley, P. D. (2006). Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, *16*, 875–884.

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*, 287–299. doi:10.1146/annurev.genet.42.110807.091442.

Holmes, J. J., Clark, S., & Modrich, P. (1990). Strand-specific mismatch correction in nuclear extracts of human and *Drosophila melanogaster* cell lines. *Proceedings of the National Academy of Sciences, USA*, *87*, 5837–5841. doi:10.1073/pnas.87.15.5837.

Hu, T. T., Eisen, M. B., Thornton, K. R., & Andolfatto, P. (2013). A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, *23*, 89–98. doi:10.1101/gr.141689.112.

Jackson, B. C., Campos, J. L., Haddrill, P. R., Charlesworth, B., & Zeng, K. (2017). Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in Drosophila. *Genome Biology and Evolution*, *9*, 102–123. doi:10.1093/gbe/evw291. arXiv:https://academic.oup.com/gbe/article-pdf/9/1/102/10240845/evw291.pdf.

Jackson, B. C., & Charlesworth, B. (2021). Evidence for a force favoring GC over AT at short intronic sites in *Drosophila simulans* and *Drosophila melanogaster*. *G3 Genes Genomes Genetics*, *11*, jkab240. doi:10.1093/g3journal/jkab240.

Johri, P., Charlesworth, B., & Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics*, *215*, 173–192. doi:10.1534/genetics.119.303002.

Kern, A., Andrew, & Begun, J., David (2005). Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: Evidence for nonequilibrium processes. *Molecular biology and evolution*, *22*, 51–62. doi:10.1093/molbev/msh269.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, *47*, 713–719.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511623486.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., & Pool, J. E. (2015). 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.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E., & Braverman, J. M. (2000). Linkage disequilibria and the site frequency spectra in the su(s) and su(wa) regions of the *Drosophila melanogaster* X chromosome. *Genetics*, *156*, 1837–1852. doi:10.1093/genetics/156.4.1837.

Lartillot, N. (2012). Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Molecular Biology and Evolution*, *30*, 356–368. doi:10.1093/molbev/mss231.

Lawrie, D. S., Messer, P. W., Hershberg, R., & Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics*, *9*, 1–18. doi:10.1371/journal.pgen.1003527.

Lucchesi, J., & Kuroda, M. (2015). Dosage compensation in *Drosophila*. *Cold Spring Harbor Perspectives in Biology*, *7*, a019398. doi:10.1101/cshperspect.a019398.

Machado, H. E., Lawrie, D. S., & Petrov, D. A. (2020). Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. *Genetics*, *214*, 511–528.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, *19*, 330–338. doi:10.1016/S0168-9525(03)00116-1.

Marais, G., Mouchiroud, D., & Duret, L. (2003). Neutral effect of recombination on base composition in *Drosophila*. *Genetical Research*, *81*, 79–87. doi:10.1017/s0016672302006079.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, *351*, 652–654.

McVean, G. A. T., & Charlesworth, B. (1999). A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetical Research*, *74*, 145–158. doi:10.1017/S0016672399003912.

Miller, D. E., Smith, C. B., Kazemi, N. Y., Cockrell, A. J., Arvanitakis, A. V., Blumenstiel, J. P., Jaspersen, S. L., & Hawley, R. S. (2016). Whole-Genome analysis of individual meiotic events in *Drosophila melanogaster* reveals that noncrossover gene conversions are insensitive to interference and the centromere effect. *Genetics*, *203*, 159–171. doi:10.1534/genetics.115.186486.

Mitchell, D., & Bridge, R. (2006). A test of Chargaff's second rule. *Biochemical and Biophysical Research Communications*, *340*, 90–94. doi:https://doi.org/10.1016/j.bbrc.2005.11.160.

Nagylaki, T. (1983). Evolution of a large population under gene conversion. *Proceedings of the National Academy of Sciences*, *80*, 5941–5945. doi:10.1073/pnas.80.19.5941.

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, USA*, *76*, 5269–5273.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., & Andolfatto, P. (2010). On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution*, *27*, 1226–34. doi:10.1093/molbev/msq046.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., & Marais, G. A. B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, *4*, 675–682. doi:10.1093/gbe/evs052.

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, *7*, e36317. doi:10.7554/eLife.36317.

de Procé, S., Zeng, K., Betancourt, A., & Charlesworth, B. (2012). Selection on codon usage and base composition in *Drosophila americana*. *Biology Letters*, *8*, 82–85. doi:10.1098/rsbl.2011.0601.

Robinson, M. C., Stone, E. A., & Singh, N. D. (2014). Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *31*, 425–433. doi:10.1093/molbev/mst220.

Rogers, R. L., Cridland, J. M., Shao, L., Hu, T. T., Andolfatto, P., & Thornton, K. R. (2014). Landscape of Standing Variation for Tandem Duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molecular Biology and Evolution*, *31*, 1750–1766. doi:10.1093/molbev/msu124.

Schrider, D., Shanku, A., & Kern, A. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*, 1207–1223. doi:10.1534/genetics.116.190223.

Singh, N. D., Arndt, P. F., Clark, A. G., & Aquadro, C. F. (2009). Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Molecular Biology and Evolution*, *26*, 1591–1605. doi:10.1093/molbev/msp071.

Singh, N. D., Arndt, P. F., & Petrov, D. A. (2005a). Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics*, *169*, 709–722. doi:10.1534/genetics.104.032250.

Singh, N. D., Davis, J. C., & Petrov, D. A. (2005b). Codon bias and noncoding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. *Journal of Molecular Evolution*, *61*, 315–324. doi:10.1007/s00239-004-0287-1.

Tachida, H. (1991). A study on a nearly neutral mutation model in finite populations. *Genetics*, *128*, 420–433.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, *123*, 585–595. arXiv:https://www.genetics.org/content/123/3/585.full.pdf.

Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y., & Thermes, C. (2004). Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Research*, *32*, 4969–4978. doi:https://doi.org/10.1093/nar/gkh823.

Vicoso, B., & Charlesworth, B. (2009). Recombination rates may affect the ratio of X to autosomal noncoding polymorphism in african populations of *Drosophila melanogaster*. *Genetics*, *181*, 1699–1701. doi:10.1534/genetics.108.098004.

Vogl, C., & Bergman, J. (2015). Inference of directional selection and mutation parameters assuming equilibrium. *Theoretical Population Biology*, *106*, 71–82.

Vogl, C., & Clemente, F. (2012). The allele-frequency spectrum in a decoupled moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theoretical Population Biology*, *81*, 197–209. doi:10.1016/j.tpb.2012.01.001.

Vogl, C., & Mikula, L. C. (2021). A nearly-neutral biallelic moran model with biased mutation and linear and quadratic selection. *Theoretical Population Biology*, *139*, 1–17. doi:10.1016/j.tpb.2021.03.003.

Vogl, C., Mikula, L. C., & Burden, C. J. (2020). Maximum likelihood estimators for scaled mutation rates in an equilibrium mutation-drift model. *Theoretical population biology*, *134*, 106–118. doi:10.1016/j.tpb.2020.06.001.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*, 431–449.

Yıldırım, B., & Vogl, C. (2023). Purifying selection against spurious splicing signals contributes to the base composition evolution of the polypyrimidine tract. *Journal of Evolutionary Biology*, *36*, 1295–1312. doi:10.1111/jeb.14205.

Zeng, K. (2010). A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. *Molecular Biology and Evolution*, *27*, 1327–1337. doi:10.1093/molbev/msq023.

Zeng, K., & Charlesworth, B. (2010). The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics*, *186*, 1411–1424. doi:10.1534/genetics.110.122150.

1. Supplement

Table S1: Mean recombination rates (in cM/Mb) of datasets binned before and after excluding telomeres and centromeres. WChr: Whole chromosome, CChr: Central regions of the chromosome arms

	Autosome		X	
	WChr	CChr	WChr	CChr
High	4.89	5.49	5.33	5.78
Modhigh	2.54	3.11	2.77	3.18
Modllow	1.48	2.15	1.71	2.25
Low	0.33	1.18	0.28	1.14

Table S2: The range of *GC* content for the *GC* binned datasets

	Autosome	X
GC1	$[0.198, 0.556]$	$[0.364, 0.615]$
GC2	$(0.556, 0.613]$	$(0.615, 0.68]$
GC3	$(0.613, 0.657]$	$(0.68, 0.731]$
GC4	$(0.657, 0.709]$	$(0.731, 0.776]$
GC5	$(0.709, 0.963]$	$(0.776, 0.944]$

Table S3: The range of *GC* content for the *GC* binned datasets *D. simulans*

	Autosome	X
GC1	$[0, 0.564]$	$[0, 0.636]$
GC2	$(0.564, 0.628]$	$(0.636, 0.698]$
GC3	$(0.628, 0.674]$	$(0.698, 0.751]$
GC4	$(0.674, 0.726]$	$(0.751, 0.799]$
GC5	$(0.726, 0.951]$	$(0.799, 1]$

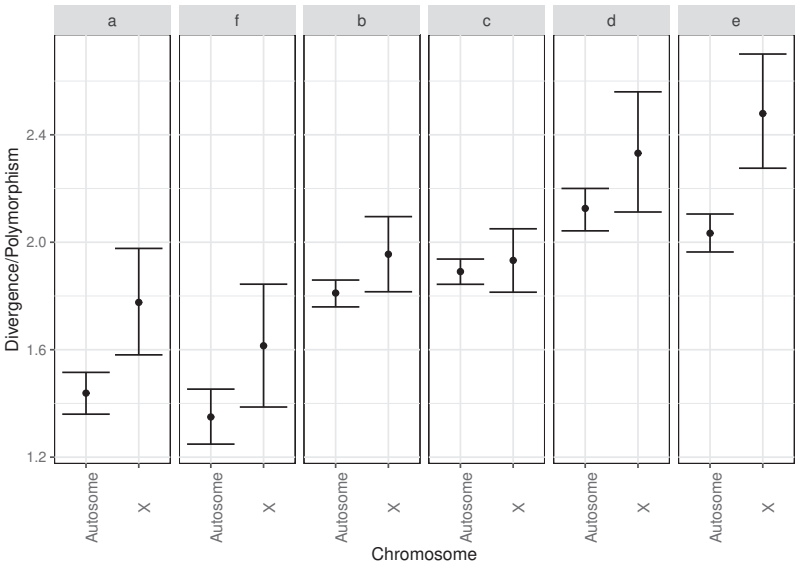

Figure S1: Divergence over polymorphism ratios for all six mutations classes from autosomes and X-chromosome, where a and f are GC-conservative, b and c GC-changing transitions and d and e GC-changing transversions.

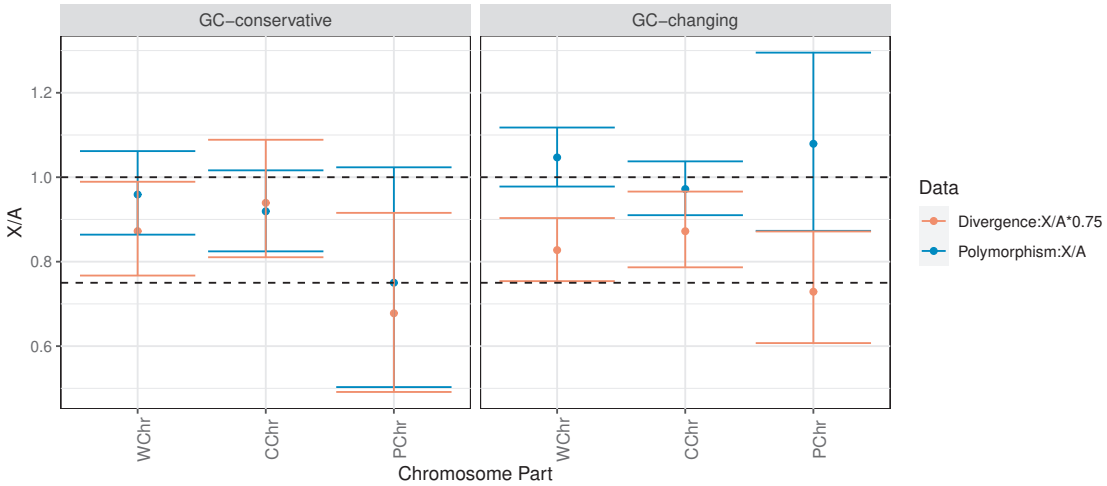

Figure S2: X/A ratios for polymorphism and divergence estimates given collectively for all GC-conservative (a and f) and changing (b, c, d, e) mutations. The divergence ratios are multiplied with 0.75; horizontal dashed lines correspond to the values of 0.75 and 1. WChr: Whole chromosome, CChr: Central regions of the chromosome arms, PChr: Peripheral regions of the chromosome arms.

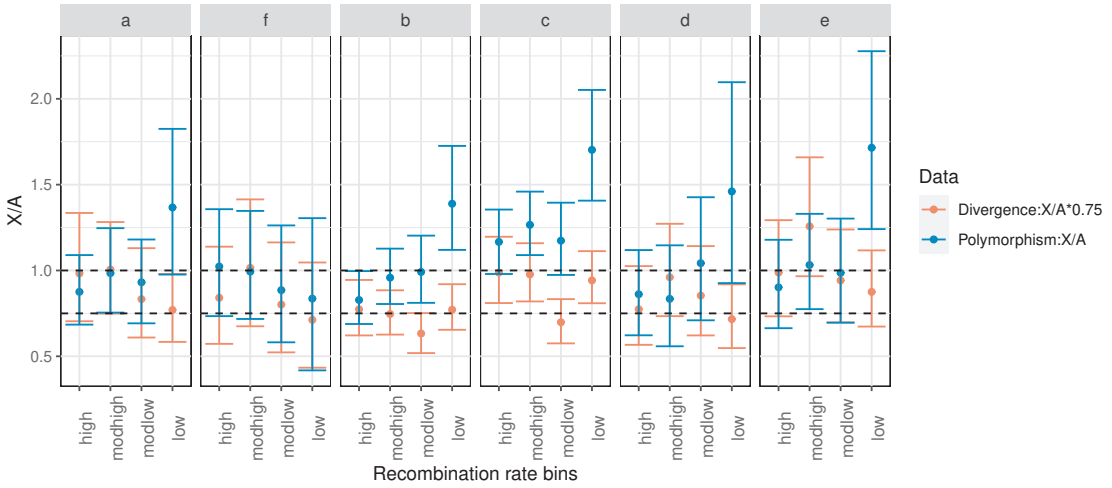

Figure S3: X/A ratios for polymorphism and divergence estimates for all six mutation classes; *a* and *f* are *GC*-conservative, *b* and *c* *GC*-changing transitions and *d* and *e* *GC*-changing transversions. Introns are binned by recombination rate using whole chromosome. The divergence ratios are multiplied with 0.75; horizontal dashed lines correspond to the values of 0.75 and 1.

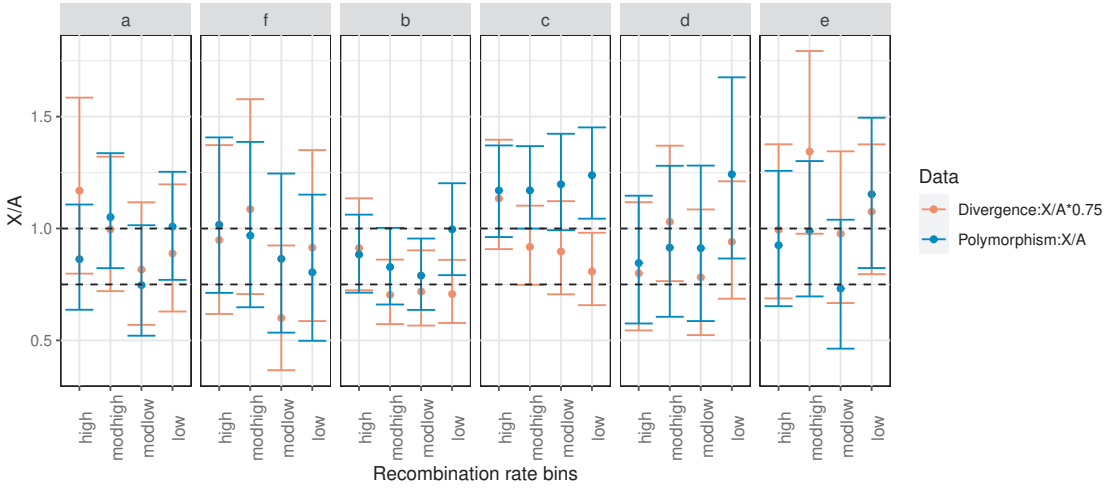

Figure S4: X/A ratios for polymorphism and divergence estimates for all six mutation classes; *a* and *f* are *GC*-conservative, *b* and *c* *GC*-changing transitions and *d* and *e* *GC*-changing transversions. Introns are binned by recombination rate after excluding telomeres and centromeres. The divergence ratios are multiplied with 0.75; horizontal dashed lines correspond to the values of 0.75 and 1.

Table S4: B values inferred from the SFS of GC-changing mutations (B_{GC}) (conditioned on the SFS of GC-conservative mutations), for autosomes, the X chromosome, and autosomes and X pooled. Estimates are either given separately for central and peripheral regions of the chromosome arms (CChr and PChr, respectively) or for whole chromosomes (WChr). All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	WChr	CChr	PChr
Autosomes	0.465	0.450	0.534
X	0.572	0.559	0.614
Pool	**0.478**	**0.464**	**0.545**

Table S5: B values inferred from the SFS of GC-changing transitions (B_{Ts}) and transversions (B_{Tv}) (conditioned on the SFS of GC-conservative mutations), for autosomes, the X chromosome, and autosomes and the X pooled. Estimates are either given separately for central and peripheral regions of the chromosome arms (CChr and PChr, respectively) or for whole chromosomes (WChr). All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	WChr		CChr		PChr	
	Transitions	Transversions	Transitions	Transversions	Transitions	Transversions
Autosome	0.421	0.557	0.395	0.570	0.557	0.487
X	0.587	0.539	0.614	0.435	0.383	1.259
Pool	**0.441**	**0.554**	**0.422**	**0.553**	**0.540**	**0.557**

Table S6: B values inferred from the SFS of all GC-changing mutations (B_{GC}), GC-changing transitions (B_{Ts}) and transversions (B_{Tv}) (conditioned on the SFS of GC-conservative mutations, with r_y) for autosomes, the X chromosome, and pooled data (autosomes+X) of *D. simulans* population. All values are significantly different from $B = 0$ ($p < 0.001$). Bold values for the pooled data indicate no significant differences between autosomes and the X.

	B_{GC}	B_{Tv}	B_{Ts}
Autosomes	0.378	0.594	0.272
X	0.311	0.480	0.232
Pool	**0.371**	**0.582**	**0.268**

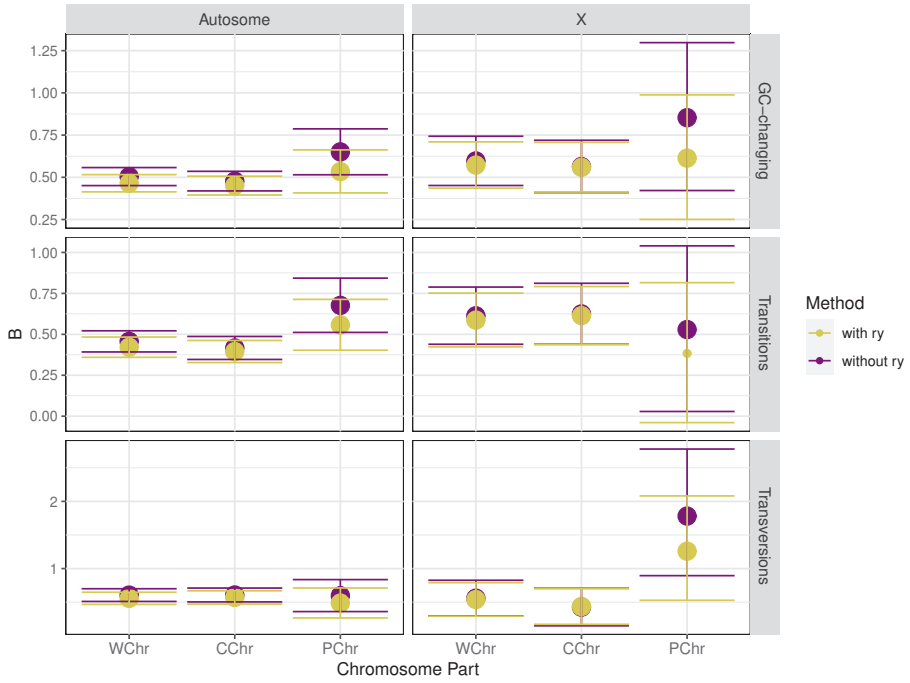

Figure S5: B values inferred from the SFS of all GC-changing mutations (B_{GC}), GC-changing transitions (B_{Ts}) and transversions (B_{Tv}) with (yellow) or without r_y (purple) for autosomes and X chromosome. The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test. Estimates are either given separately for central and peripheral regions of the chromosome arms (CChr and PChr, respectively) or for whole chromosomes (WChr).

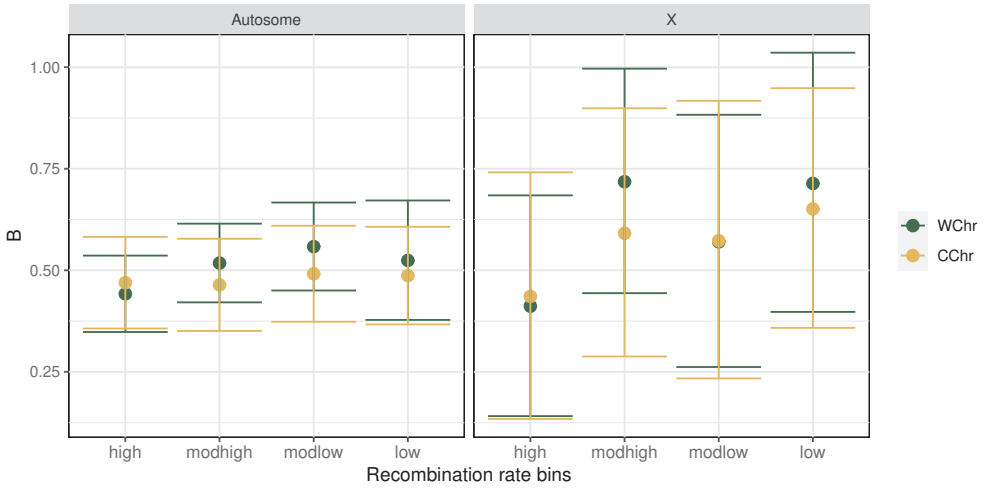

Figure S6: B values inferred from the SFS of GC-changing mutations (B_{GC}) for autosomes and X chromosomes. Introns are binned by recombination rate before (green, WChr: Whole chromosome) and after excluding telomeres and centromeres (yellow, CChr: Central regions of the chromosome arms). Confidence intervals are constructed from likelihood ratio test.

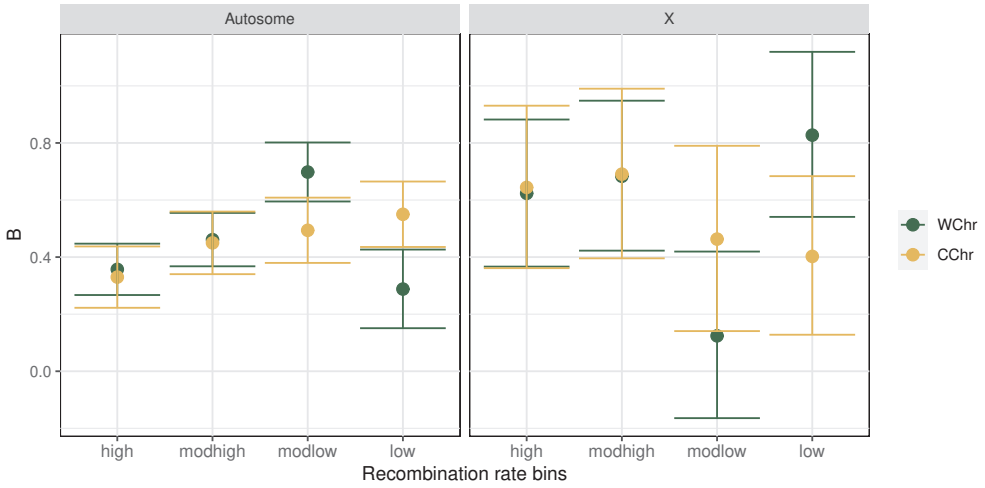

Figure S7: B values inferred from the SFS of GC-changing mutations (B_{GC}) (conditioned on the SFS of GC-conservative mutations) for autosomes and X chromosomes. Introns are binned by recombination rate before (green, WChr: Whole chromosome) and after excluding telomeres and centromeres (yellow, CChr: Central regions of the chromosome arms). Confidence intervals are constructed from likelihood ratio test.

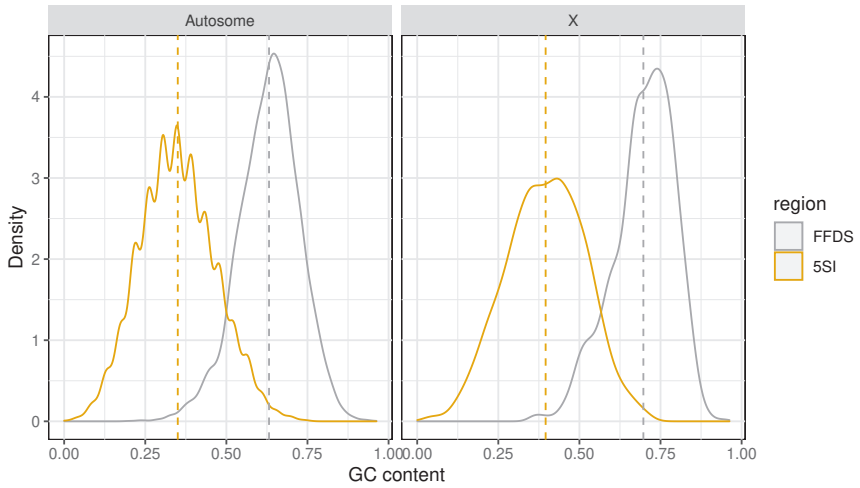

Figure S8: The mean *GC* content distributions for FFDS (gray) and 5SI (yellow) of autosomes and the X chromosome. The vertical dashed lines corresponds to the means of the distributions.

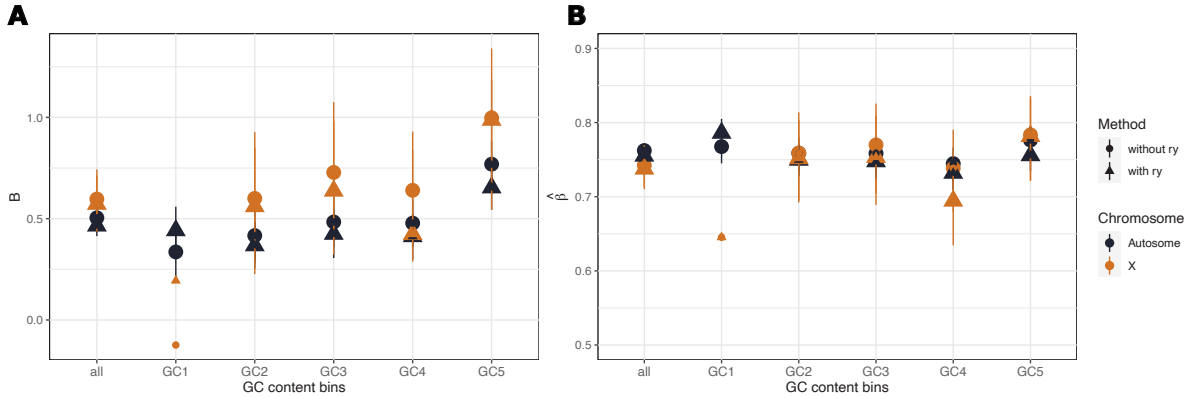

Figure S9: Results from *GC* content binned data of *D. melanogaster* population with (triangle) and without (round) parameter r_y to account for demography. A) B_{GC} values inferred from the SFS of *GC*-changing mutations for autosomes (black) and X chromosome (brown). The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test. B) Mutation bias estimated conditional on B for for autosomes (black) and X chromosome (brown). Estimates are given for all introns and for introns binned by the mean *GC* content of the FFDS of the same genes. *GC* content increases from GC1 to GC5 and the ranges are given in S2

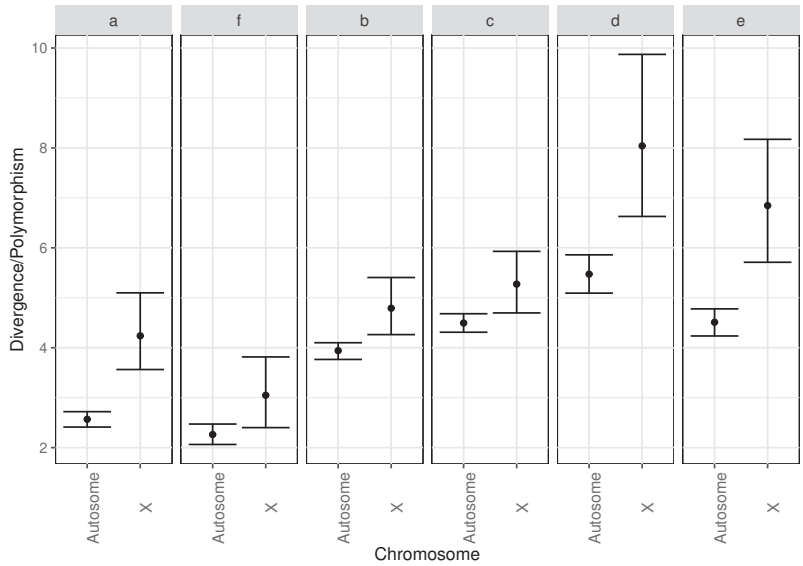

Figure S10: Divergence over polymorphism ratios for all six mutation classes for autosomes and the X chromosome of *D. simulans*, where *a* and *f* are *GC*-conservative, *b* and *c* *GC*-changing transitions and *d* and *e* *GC*-changing transversions.

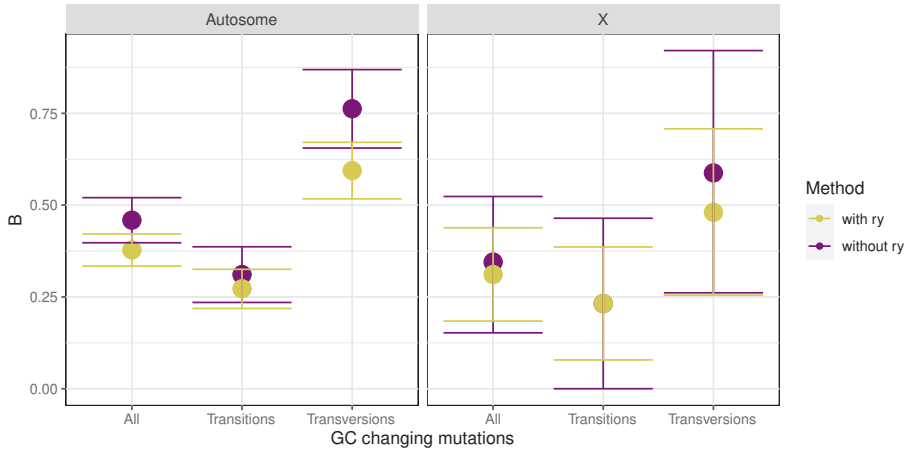

Figure S11: B values inferred from the SFS of all *GC*-changing mutations (B_{GC}), *GC*-changing transitions (B_{Ts}) and transversions (B_{Tv}) with (yellow) or without r_y (purple) for autosomes and X chromosome of *D. simulans* population. The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test.

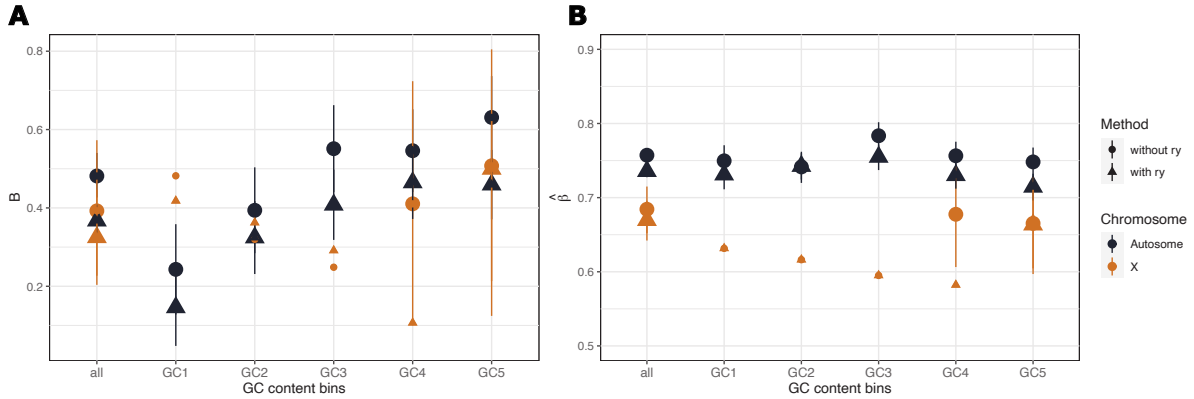

Figure S12: Results from GC content binned data of *D. simulans* population with (triangle) and without (round) parameter r_y to account for demography. A) B_{GC} values inferred from the SFS of GC-changing mutations for autosomes (black) and X chromosome (brown). The big dots represents the estimates significantly different from $B = 0$. Confidence intervals are constructed from likelihood ratio test. B) Mutation bias estimated conditional on B for for autosomes (black) and X chromosome (brown). Estimates are given for all introns and for introns binned by the mean GC content of the FFDS of the same genes. GC content increases from GC1 to GC5 and the ranges are given in S3