Evidence for a force favouring GC over AT at short intronic sites in *Drosophila simulans* and *D. melanogaster*

Ben Jackson^{†*} and Brian Charlesworth[†]

[†]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

*Corresponding Author: ben.jackson@ed.ac.uk

12 **Running Title**

13 GC-biased gene conversion in *Drosophila*

14

15 **Key Words**

16 Evolution, GC-biased gene conversion, Mutation, *Drosophila*

17

18

Abstract

Population genetics studies often make use of a class of nucleotide site free from selective pressures in order to make inferences about population size changes or natural selection at other sites. If such neutral sites can be identified, they offer the opportunity to avoid any confounding effects of selection. Here we investigate evolution at putatively neutrally evolving short intronic sites in natural populations of *Drosophila melanogaster* and *D. simulans*, in order to understand the properties of spontaneous mutations and the extent of GC-biased gene conversion in these species. Use of data on the genetics of natural populations is advantageous because it integrates information from large numbers of individuals over long timescales. In agreement with direct evidence from observations of spontaneous mutations in *Drosophila*, we find a bias in the spectrum of mutations towards AT basepairs. In addition, we find that this bias is stronger in the *D. melanogaster* lineage than the *D. simulans* lineage. The evidence for GC-biased gene conversion in *Drosophila* has been equivocal. Here we provide evidence for a weak force favouring GC in both species, which is stronger in *D. simulans*. Some homologous short intronic sites have diverged in GC content between the two species, which may have been caused by lineage-specific changes in the extent to which different regions of the genome are subject to a GC (or AT) -favouring force.

Introduction

Population genetics studies often make use of a class of nucleotide site that is considered to be free from selective pressures, for the purpose of making inferences about the demographic

history of a population (e.g., Fagundes et al. 2007; Obbard et al. 2012; Garud et al. 2015). Such sites are also used as a neutral comparator for the purpose of estimating the parameters of natural selection acting on other types of sites, while controlling for other processes that affect the allelic composition of populations, such as population size changes, genetic drift and mutation. For example, the McDonald-Kreitman test and its extensions (McDonald and Kreitman 1991; Fay et al. 2001; Smith and Eyre-Walker 2002; Welch 2006; Messer and Petrov 2013) rely on a class of site that are assumed to evolve neutrally, whose relative levels of between-species divergence and within-species variability are contrasted to those for a putatively selected class, generating an estimate of the fraction of substitutions that have been fixed by positive selection as opposed to drift. More recent methods for inferring the distribution of fitness effects of new mutations also rely on a neutrally evolving class of site, especially to correct for the effects of past population changes (Keightley and Eyre-Walker 2007; Boyko et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011; Galtier 2016; Kim et al. 2017; Tataru et al. 2017).

In a number of species of *Drosophila*, there is evidence for the functional significance of a large fraction of the genome, including the action of both purifying and positive selection on intronic and intergenic sites (Andolfatto 2005; Haddrill et al. 2005; Halligan and Keightley 2006; Haddrill et al. 2008; de Procé et al. 2012). Synonymous changes may also be subject to weak selection for preferred codons, which can affect allele frequencies in species with sufficiently large population sizes for such selection to be effective, including *Drosophila melanogaster* and *D. simulans* (Begun 2001; Vicario et al. 2007; Jackson et al. 2017; Machado et al. 2020).

A candidate for a class of neutral nucleotide site in *Drosophila* is the 8-30 basepair region from the 5' end of introns shorter than 66bp, but ≥ 23 bp after removing splice sites (referred to here as SI sites), which have been shown to have the highest between-species divergence and within-population diversity compared to other regions of the genome (Halligan and Keightley 2006; Parsch et al. 2010). These patterns are suggestive of a low level of purifying selection, and consequently a lack of functional importance. This makes them a good candidate for a neutral comparator of the type required by the methods mentioned above. For example, short introns have been used to infer strong purifying selection at 4-fold degenerate sites in *D. melanogaster* (Lawrie et al. 2013; Machado et al. 2020), to fit demographic models to North American *D. melanogaster* in order to determine appropriate parameters for inferring selection from haplotype statistics (Garud et al. 2015), and to quantify population structure in European *D. melanogaster* (Kapun et al. 2020). Sites outside the 8-30bp region but within short introns are probably more constrained because they are functionally important for mRNA splicing (Green 1986; Mount et al. 1992; Kennedy and Berget 1997; Halligan and Keightley 2006).

If SIs do indeed evolve in the absence of selective constraint, they provide an opportunity to investigate processes that affect the composition of genomes other than natural selection. Subsequent studies of evolution at SI sites in *Drosophila* have provided evidence for context-dependent mutational patterns (Clemente and Vogl 2012) and the possible action of GC-biased gene conversion (gBGC) (Jackson et al. 2017). Evidence for gBGC in *Drosophila* genomes is equivocal, with some suggestion that it operates on the X chromosome in *D. simulans* (Haddrill and Charlesworth 2008) and *D. americana* (de Procé et al. 2012), and on both autosomes and the X chromosomes in *D. simulans* and *D. melanogaster* (Jackson et al.

2017), while other studies have found little or no evidence for it (Clemente and Vogl 2012; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014).

Direct observations of spontaneous mutations in *D. melanogaster*, as well as analyses of rare segregating polymorphisms, show that mutation is biased towards GC to AT basepair substitutions (Assaf et al. 2017), and population genetic studies have suggested that the extent of this bias has increased at some point in the evolutionary past (Kern and Begun 2005; Zeng and Charlesworth 2010; Clemente and Vogl 2012). Laboratory studies of mutation are limited in power because mutations are rare – with a mutation rate of 5×10^{-9} per basepair (Assaf et al. 2017) we expect 0.7 mutations per haploid genome per generation in a genome containing 140 million basepairs. But an examination of the population genetics of natural populations provides the opportunity to integrate evidence from over long evolutionary timescales. In combination with information from sections of the genome that are free from directional evolutionary processes, this procedure should provide further insights into the nature of spontaneous mutations in *Drosophila* and the forces affecting mutations with little or no effects on fitness.

To investigate possible non-selective directional evolutionary processes in *Drosophila*, we have investigated evolution at short intron (SI) sites, using polymorphism data from populations from the putative ancestral ranges of *D. melanogaster* and *D. simulans*, as well as data on between-species divergence. Our study refines the analyses of Jackson *et al.* (2017), which primarily focussed on four-fold degenerate sites, since their analyses of SI sites were hampered by insufficient amounts of data and a poorer quality annotation of the *D. simulans* genome than the one used here. We show that a subset of SI sites are probably subject to directional evolutionary pressures, with GC alleles being favoured over AT alleles at SI sites

with the highest GC contents, suggesting the action of GC-biased gene conversion in both species. This has implications for the use of short introns as a neutrally-evolving reference in population genetics, and also sheds light on the dynamics of genome evolution in *Drosophila*. The study also provides further evidence for the existence of a strong GC to AT mutational bias in *Drosophila*. Its magnitude appears to be independent of the GC content of short introns, and has apparently increased along the *D. melanogaster* lineage following its divergence from the common ancestor of *D. melanogaster* and *D. simulans*.

Materials and Methods

Sequence data from *D. simulans* and *D. melanogaster*

We have analysed a previously published population sample of 21 lines of *Drosophila simulans*, derived from the putatively ancestral Madagascan population (the MD lines of Jackson et al. 2017). The sampling, maintenance, sequencing and variant-calling procedures for these lines were fully described in Rogers *et al.* (2014), Jackson *et al.* (2017) and Becher *et al.* (2020). Briefly, publicly available raw read data in FASTQ format for these 21 lines were downloaded from the European Nucleotide Archive and mapped to version 2.02 of the *D. simulans* reference genome (FlyBase release 2017_04) using BWA MEM (Li and Durbin 2009). We sorted, merged and marked duplicates on the resulting BAM files using Picard Tools version 2.8.3 (<https://broadinstitute.github.io/picard/>). Variants were called for each line individually using the HaplotypeCaller tool from GATK version 3.7 (McKenna et al. 2010) with the options `–emitRefConfidence BP_RESOLUTION` and `–max-alternate-alleles` 2. VCF files containing all 21 lines were generated from the output of HaplotypeCaller using

the GATK tool GenotypeGVCFs. We treated sites that remained heterozygous within samples after inbreeding as follows: at each heterozygous site within a sample, one allele was chosen as the haploid genotype call at that site with a probability proportional to its coverage in the sample. The alternative allele was discarded (Jackson et al. 2017).

We also downloaded publicly available sequence data for 197 lines of *D. melanogaster* sampled from Zambia (ZI lines) from the Drosophila Genome Nexus (DGN) (<https://www.johnpool.net/genomes.html>), and converted these data to FASTA format using a custom shell script. Using the information reported in the supplement to Lack et al. (2015) we retained 69 ZI lines that showed no evidence of admixture with European populations. The sampling, sequencing and variant-calling procedures, and the procedure for defining admixture tracts for these ZI lines were described fully in Pool et al. (2012) and Lack et al. (2015). The ZI sample is maximally diverse and minimally affected by cosmopolitan admixture among the populations in the DGN (Lack et al. 2015), and also provides the largest sample of African *D. melanogaster* genomes.

Between-species alignments

We used the multispecies alignment between *D. melanogaster*, *D. simulans* and *D. yakuba* from Zeng et al. (2019). Briefly, multispecies alignment was performed between the reference genomes of *D. simulans* (v2.02), *D. melanogaster* (v5.57), and *D. yakuba* (v1.05) using the MULTI-Z pipeline described by Barton and Zeng (2018). Reference genomes were downloaded from FlyBase, and repeat regions were soft-masked using RepeatMasker (<http://www.repeatmasker.org/>) with the default database for *Drosophila*. Pairwise alignments were generated between *D. melanogaster* and *D. simulans*, and between *D.*

melanogaster and *D. yakuba* using LASTZ (Harris 2007), which were chained and netted using axtChain and chainNet (Kent et al. 2003). Single coverage was generated using single_cov2.v11 from the MULTIZ package (Blanchette et al. 2004) and the pairwise alignments were aligned with MULTIZ to create a three-way multiple alignment.

Defining short intronic sites

To define short intronic (SI) sites, we first carried out the following procedure separately for each of *D. melanogaster* and *D. simulans*. We used the information in the header lines of the FlyBase FASTA file of introns for version 2.02 (5.57) of the *D. simulans* (*D. melanogaster*) reference genome to extract coordinates of the 8–30 bp region of introns that were ≤ 65 bp in length, after checking that this region did not overlap with an exon, an intron of length more than 65 bp, or the non-8–30 bp portion of an intron of length ≤ 65 bp, using information from the gff format annotation of the *D. simulans* (*D. melanogaster*) reference genome version 2.02 (5.57) downloaded from FlyBase.

Using the resulting short intronic positions in each species and the whole genome alignment described above, we defined a set of homologous sites that were annotated as short intronic sites in both *D. simulans* and *D. melanogaster*, using the script non_ref_intersect.py from the WGAbed package (<https://henryjuho.github.io/WGAbed/>) and the bedtools subroutine intersectBed (Quinlan and Hall 2010), as well as additional custom shell and Python scripts. We generated an alignment for each *D. melanogaster* short intron region that overlapped with a *D. simulans* short intron region, yielding polymorphism data for the ZI and MD lines, as well as the corresponding sequences from each of the *D. melanogaster* v5.57, *D. simulans* v2.02 and *D. yakuba* v1.3 reference sequences.

192

193 At this stage, we retained SI sites only if the following additional conditions were met: they
194 were located on an autosome in both *D. melanogaster* and *D. simulans*; there were no
195 missing alleles in any of the three reference sequences; they were not soft-masked as
196 repetitive in any of the three reference sequences; they did not overlap with an indel in the *D.*
197 *simulans* variant callset; QUAL $\geq$ 30 in the *D. simulans* variant callset; they did not lie in a
198 non-crossover region in either the *D. melanogaster* genome (as defined in Campos et al.
199 2012) or in the *D. simulans* genome (as defined in Becher et al. 2020). In total, we retained
200 167,147 autosomal SI sites for divergence-based analyses. For polymorphism-based analyses,
201 such as using derived allele frequencies or site frequency spectra (see below), we further
202 excluded sites with any missing polymorphism data in the population under consideration.
203 We retained 163,998 and 145,747 sites for polymorphism-based analyses of the MD lines and
204 the 69 ZI lines, respectively.

205

206 For the purposes of comparing regions with different GC contents, we assigned SI sites to
207 bins of equal numbers of introns, according to their GC content. We calculated the GC
208 contents of SI sites from the full 8-30bp region of the short intron under consideration in the
209 reference sequences for *D. melanogaster* and *D. simulans*. Then we assigned SI sites to bins
210 in three different ways. Firstly, we took the mean of the GC content values for each
211 homologous pair of introns and applied this single value when grouping both species' SIs.
212 This results in the same set of homologous sites in each bin for analyses of divergence and
213 polymorphism in both lineages. Secondly, we grouped introns into species-specific bins, for
214 *D. melanogaster* by using the GC content calculated from the *D. melanogaster* reference
215 sequence, and for *D. simulans* by using the GC content calculated from the *D. simulans*
216 reference sequence. This means that homologous sites may be assigned to different bins in

analyses of the *D. simulans* lineage from those in analyses of the *D. melanogaster* lineage, but ensures a perfect relationship between within-species GC content and bin. Below, we refer to these two binning strategies as ‘mean’ and ‘species’, respectively. The two GC contents are closely correlated, as is expected given the slow evolution of GC content (Supplementary Figure S1), but their grouping of SI sites sometimes resulted in contrasting conclusions about patterns of evolution, which we discuss below. Thirdly, we grouped SI sites into bins by the difference in GC content between *D. melanogaster* and *D. simulans* per short intron. We argue below that binning by the mean criterion for GC content is appropriate for analyses of substitution patterns, whereas binning by the species criterion is more appropriate for polymorphism analyses. Binning by the between-species difference in GC content sheds light on processes in regions of the genome that have diverged in GC content between the two species.

To obtain confidence intervals (CIs) around point estimates of statistics, we bootstrapped by sampling introns with replacement 1000 times until the bootstrapped sample was the same size as the observed sample. For each bootstrap sample, we recalculated the statistic of interest. We used the 2.5% and 97.5% quantiles of the resulting distribution as the upper and lower bounds of the 95% CI (Efron 1979).

Analyses of between-species divergence

We used the GTR-NHb model of base substitution modified to generate sub-optimal ancestral sequences, as implemented in the baseml program of PAML version 4.8 (Yang 2007), in order to reconstruct the base content of the *melanogaster-simulans* (*ms*) ancestor,

and counted substitutions along lineages according to the Expected Markov Counting method of Matsumoto *et al.* (2015). This method should be more accurate than maximum parsimony or use of a single best reconstruction under complex patterns of base substitution, which are likely apply to *Drosophila* (Matsumoto et al. 2015). We checked that the GTR-NHb fitted the data better than the stationary GTR model, also implemented in PAML, using likelihood ratio tests – this was true in all cases. For each bin, we ran baseml ten times and manually checked for convergence by examining the likelihood output of the model. In the results presented below, we refer to G and C alleles as strong (S) and to A and T alleles as weak (W). We categorised the number of substitutions from the *ms* ancestor to the extant *simulans* sequence or to the extant *melanogaster* sequence into the following classes: the total number of substitutions from strong to weak alleles, $N_{S>W}$; the total number of substitutions from weak to strong alleles, $N_{W>S}$; and the total number of substitutions from strong to strong alleles or from weak to weak alleles, N_{neu} . We denote the number of GC sites in the ancestral sequence by L_{GC} and the number of AT sites in the ancestral sequence by L_{AT} . We define the substitution rate from strong to weak alleles as $r_{S>W} = N_{S>W} / L_{GC}$, and the substitution rate from weak to strong alleles as $r_{W>S} = N_{W>S} / L_{AT}$. We obtained the expected numbers of substitutions and the predicted ancestral base content by parsing the output of PAML, using custom scripts in R (R Core Team 2018).

Analyses of polymorphism data

We divided SI sites into the same sets of five bins as used for the divergence-based analyses. For each population, we excluded sites with missing data in the polymorphism sample as well as segregating sites of more than two alleles, and then used est-sfs v2.03 (Keightley and

Jackson 2018) to calculate the probability of the major allele being ancestral for each segregating site. We used the Kimura 2-parameter model of base substitution and two outgroups (*D. yakuba* and one of either *D. melanogaster* or *D. simulans*, depending on the species to which the polymorphism data belonged) to run est-sfs, and carried out 10 maximum likelihood searches for each bin to check for convergence. Using the results, we constructed separate unfolded site-frequency spectra (SFSs) for segregating $S > W$, $W > S$ and neutral ($W = S$ or $S = S$) mutations. We used these SFSs to calculate the mean derived allele frequency (DAF) for each class of change, and as an input for the method of Glémin *et al.* (2015) for estimating the mutation and selection parameters.

This method uses the three unfolded SFSs for segregating sites described above to estimate γ and κ , where $\gamma = 4N_e s$ is the scaled strength of selection for GC (S) alleles, and κ is the mutational bias parameter u/v . Here, s is the selection coefficient against heterozygotes for W and S alleles (semi-dominance is assumed), u is the mutation rate from S to W, and v is the mutation rate from W to S. This method uses the three unfolded SFSs described above to disentangle γ and κ . The method is capable of taking into account polarization errors, which can lead to upwardly biased estimates of γ (Hernandez et al. 2007), by incorporating them into the model and estimating them jointly with the parameters of interest. It is also capable of correcting for demographic effects, by introducing nuisance parameters to correct for distortions in the SFS due to demography (following Eyre-Walker et al. 2006). We estimated γ and κ using the R code provided in the supplement of Glémin et al. (2015). We refer to the models using this method with the same notation as in Glémin et al. (2015). These are model M0, with $\gamma = 0$, and no correction for polarization errors; M1, with $\gamma \neq 0$ and no correction for polarization errors; and M0* and M1*, which are the equivalent models after correcting for polarization errors. We compared the different models using likelihood ratio tests.

292

293 **Computational methods**

294

295 This work made use of GNU parallel (Tange 2011).

296

297

298 **Data Availability**

299

300 All the code required to replicate the analyses presented here is available on Github

301 (https://github.com/benjaminjcjackson/dros_gBGC), including the final dataset of

302 homologous short intronic sites in plain text format.

303

304

305 **Results**

306

307 **Summary of polymorphism and divergence results**

308

309 In agreement with previous analyses (Jackson et al. 2017), the population of *D. simulans*

310 from its putatively ancestral range in Madagascar is more diverse than either population of *D.*

311 *melanogaster* – nucleotide diversity (π) at SI sites is 0.037 for the MD sample and 0.016 for

312 the ZI sample (Table 1). The site frequency spectrum is more skewed towards rare variants in

313 *D. simulans* than in *D. melanogaster*, as shown by the larger absolute values of Tajima's *D*

314 and the proportion of singletons in the MD sample, compared to ZI (Table 1). The ratios of

315 the proportions of singletons to their expected values under neutrality are 1.83 and 1.48 for

316 MD and ZI, respectively. Parsch et al. (2010) reported the divergence between *D.*

melanogaster and *D. simulans* at the 8-30bp region of introns < 66bp long to be 0.123, with the divergence along the *D. melanogaster* lineage equal to 0.064. These values are nearly identical to those for our dataset with all sites concatenated: 0.124 (0.121-0.126) for the *D. melanogaster* and *D. simulans* net divergence, 0.0655 (0.0637-0.0674) for the *D. melanogaster* branch, and 0.0582 (0.0567, 0.0598) for the *D. simulans* branch (the brackets indicate the confidence intervals for the means).

Testing for fixation bias

If base composition is at statistical equilibrium under mutation, drift and selection over the period of time covered by an evolutionary lineage, there should be equal numbers of substitutions from *S* (G or C) to *W* (A or T) alleles and from *W* to *S* alleles (Bulmer 1991; Akashi 1995; Charlesworth and Charlesworth 2010 p.272). Using the same methodology as here, we previously reported a slight overall AT-bias in substitutions in *D. simulans* autosomal short introns ($\chi^2 = 5.55$, $df = 1$, $P = 0.019$) (see Table 2 in Jackson et al. 2017). The results obtained from our new dataset suggests an equilibrium base composition in *D. simulans* autosomal SIs when we concatenate all SI sites. $N_{W>S}$ for all SI sites combined in *D. simulans* was 3500, and $N_{S>W}$ was 3584, which do not differ significantly from a ratio of 1:1 ($\chi^2 = 0.99$, $df = 1$, $P = 0.32$).

After binning SIs by their mean GC content across species, there was no obvious relationship between GC content and the ratio $N_{W>S}/N_{S>W}$. The 95% confidence intervals obtained by bootstrapping overlap unity for all bins (Figure 1A). This suggests that aggregating sites with different GC contents does not mask any substitution patterns that are specific to sequence context. In agreement with previous results (Akashi et al. 2006; Jackson et al. 2017), *D.*

melanogaster shows an overall bias towards AT-fixation for all sites combined ($N_{W>S} = 3478$, $N_{S>W} = 4760$, $\chi^2 = 199$, $df = 1$, $P < 0.001$). When *D. melanogaster* SI sites are binned by mean GC content, all of the bins exhibit an AT fixation-bias, which increases with increasing GC content (Figure 1B). This implies that *D. melanogaster* has experienced a change in the forces acting on GC content, such that its GC content is currently not at equilibrium; this could either be a change in mutational bias towards GC>AT mutations, or a reduction in the scaled intensity of a selective force or biased gene conversion favouring GC, possibly reflecting a reduction in effective population size. In contrast, *D. simulans* appears to be approximately in equilibrium.

On the other hand, when we used the species-specific GC content for binning SIs, in both *D. simulans* and *D. melanogaster* the ratio $N_{W>S}/N_{S>W}$ increases with increasing GC content (Figure 1C, D) – i.e. there have been proportionally more changes from AT to GC than from GC to AT in more GC-rich regions. In *D. simulans*, only the middle bin appears to be at equilibrium according to this analysis. The lowest two GC content bins exhibit an excess of $S>W$ substitutions, and the highest two GC content bins exhibit an excess of $W>S$ substitutions (Figure 1C). In *D. melanogaster*, the bottom four GC bins exhibit an excess of $S>W$ substitutions, and the highest GC content bin might be at equilibrium (Figure 1D).

There are, however, strong reasons for thinking that this method of binning produces biases in inferences concerning the relationship between substitution patterns and GC content.

Consider, for example, the bin with the lowest GC content in a given species. With a substitution rate of around 0.06 along its lineage back to its common ancestor with the other species, the expected number of changes within an SI along the lineage is of the order of $0.06 \times 23 = 1.4$, or less. The chance that both of these substitutions are both $W>S$ is thus very high. If SIs that have the lowest GC content are chosen according to the species-specific GC

content, these are automatically enriched for excess of $W>S$ changes as opposed to $S>W$ changes. The converse applies to SIs chosen for a low GC content. No such selection bias is expected if SIs are chosen on the basis of the mean of their GC contents for the two species. For this reason, the estimates based on the mean bins are more likely to represent the underlying evolutionary process.

Analyses of polymorphism data

The polymorphism data were used to investigate the parameters of mutation and selection acting on GC versus AT variants. In this case, binning SIs by the species-specific GC content is probably more appropriate than using the mean values across species, since it provides a better reflection of the sequence composition over the comparatively short time-scale experienced by currently segregating variants.

We examined the nature of the forces acting on polymorphic variants in two ways. First, for each bin, we calculated the mean derived variant frequency for different classes of mutations at segregating sites. These classes involved either GC to AT variants ($DAF_{S>W}$), AT to GC ($DAF_{W>S}$), or GC to CG or AT to TA (DAF_{neu}). These statistics should shed light on the processes of interest in *Drosophila* genome evolution as follows. If the mutational process has shifted towards a greater GG>AT bias within the last $4N_e$ generations (the time frame relevant to polymorphism data), as has previously been suggested for *D. melanogaster* (Akashi et al. 2006; Jackson et al. 2017), we expect $DAF_{S>W} < DAF_{W>S}$, because mutations from GC to AT should be younger on average, even under neutrality. Furthermore, if such a change in mutational bias were genome wide, we do not expect a relationship between DAF and GC content. In contrast, if gBGC or a selective force favouring GC is in operation, we

expect to have $DAF_{W>S} > DAF_{neu} > DAF_{S>W}$ (Jackson et al. 2017). In addition, if the strength of such a force varies across the genome and has influenced its GC content, we expect this relationship be stronger in regions of higher GC content – that is, $DAF_{S>W}$ should be negatively related to GC content and $DAF_{W>S}$ should be positively related to GC content. This second pattern, suggestive of a force favouring GC, is indeed what we observe in *D. simulans*, for both the mean and species-specific method of binning SIs by GC content, with a clearer pattern for the species-specific bins than the mean bins (Figure 2A and 2C).

For the Zambian sample of *D. melanogaster*, we observe the pattern supporting a GC favouring force in the top three highest GC content bins when we bin introns using the species-specific method (Figure 2D), but the relationships are less clear when the mean bins are used (Figure 2B). In the case of the mean bins, the three highest GC content bins exhibit $DAF_{W>S} > DAF_{S>W}$, although $DAF_{S>W}$ seems to decrease with decreasing GC content across all five bins, there is no obvious relationship between $DAF_{W>S}$ and GC content (Figure 2B).

One possible explanation is that these differences between the mean and species bins reflect the fact that the GC content of the mean bins reflects the ancient GC contents of the SIs involved. If, for example, there has been a shift towards a weaker force favouring GC in both lineages, but whose strength is nevertheless still correlated with GC content (as indicated by the analysis of γ below), the mean bins will on average be associated with the past GC content of the bins, so that there will be a less clear relationship of their derived variant frequencies to GC content than for the species bins. The polymorphism results for the species bins would then provide a more reliable picture than those from the mean bins.

Overall, these results suggest the presence of a GC-favouring force in *D. simulans* and, probably to a lesser extent, in *D. melanogaster*. In order to quantify this force, we used the method of Glémin *et al.* (2015) to calculate $\gamma = 4N_e s$, the scaled strength of selection or biased gene conversion favouring GC alleles. In no cases did models correcting for polarisation errors fit the data better than the equivalent model without corrections. This may be because the method we used to polarise segregating sites is less prone to mis-inference than methods such as maximum parsimony (Keightley and Jackson 2018). Both sets of models returned very similar values of γ within each species (Figure 3). Below, we report the results from models M0 and M1, which do not correct for polarization error (see Methods).

In *D. simulans*, there is little evidence for a force favouring GC alleles in the lowest GC content bin, with the exception of model M1 in the lowest mean bin, which yields an estimate of $\gamma = 0.38$ (Figure 3A). In the remaining four bins, γ is significantly greater than zero, and increases monotonically from 0.69 in the second lowest GC bin to 1.56 in the highest GC bin for the mean bins (Figure 3A). For the species bins, the relationship between GC content and gamma is somewhat more pronounced, with $\gamma = 0.60$ in the second lowest GC bin, rising to $\gamma = 2.10$ in the highest GC content bin (Figure 3C).

For the mean bins in *D. melanogaster*, the pattern is similar, but the force favouring GC is weaker. Neither of the lowest two GC bins show evidence that GC is favoured (Figure 3B). The top three GC bins show evidence for a force favouring GC of a similar magnitude to each other (0.68, 0.93 and 0.86 for bins 3, 4 and 5, respectively). For the species bins, the lowest bin shows evidence for a force favouring AT, with $\gamma = -0.79$ (Figure 3D). The second lowest GC bin shows no evidence that either strong or weak alleles are preferred, and the top three bins show evidence for a force favouring GC. As was found for *D. simulans*, the

relationship between GC content and γ is more pronounced for the species bins than the mean bins. Similar to the interpretations discussed for the DAF patterns, it seems plausible that the action of gBGC or selection has diverged somewhat between the two lineages, and that grouping introns by their mean GC content across the lineage dampens the signal of this effect.

This method also allows the estimation of κ , the mutational bias parameter (Figure 4), which is estimated jointly with γ (Glémin et al. 2015). We report values of κ from model M1. In *D. melanogaster*, κ seems fairly insensitive to GC content. From the mean bins, values of κ for bins 1 to 5 are 3.22, 3.15, 3.14, 3.13 and 2.67, respectively (Figure 4B). From the species bins, the values of κ are 3.09, 3.13, 3.30, 3.10 and 2.77 (Figure 4D). These values are close to the estimate derived from a meta-analysis of direct observations of spontaneous mutations in *D. melanogaster* mutation accumulation experiments, which was 3.35 (95% CIs: 3.00-3.76) (Assaf et al. 2017). In *D. simulans*, κ is somewhat lower, and seems to be slightly negatively correlated with GC content. Estimates for the mean bins are 2.81, 2.56, 2.57, 2.48 and 2.38 for bins 1 to 5, respectively (Figure 4), and estimates for the species bins are 2.82, 2.61, 2.50, 2.53 and 2.40 (Figure 4C). This is in agreement with the hypothesis of an increase in the GC to AT mutational bias in the *D. melanogaster* lineage relative to the *D. simulans* lineage, which has been proposed before (Takano-Shimizu 2001; Kern and Begun 2005; Zeng and Charlesworth 2010; Clemente and Vogl 2012).

Patterns of substitution and their relationship with GC content

If a fraction of SI sites is subject to a weak force favouring GC over AT, this should be reflected in the patterns of substitution rates. As a null hypothesis, we might expect sites in

the lowest GC content bins, where there is little evidence for a force favouring GC, to exhibit substitution rates that reflect the mutational bias inferred from polymorphism data and mutation accumulation experiments – under neutrality, substitution rates and mutation rates are equal (Wright 1938; Kimura 1968). For higher GC content bins, where there is evidence for an advantage to GC, we expect a higher rate of substitution of GC alleles relative to the lower GC content bins. To investigate this, we reconstructed ancestral states using PAML (see Materials and Methods for details), and counted the numbers of $S>W$ and $W>S$ substitutions along each lineage, in order to estimate the ratio of the two substitution rates, $R = r_{S>W}/r_{W>S}$, for both the *D. melanogaster* and *D. simulans* lineages. Note that $R = \kappa$ under strict neutrality.

In agreement with this hypothesis, R and GC content are negatively correlated in both *D. simulans* and *D. melanogaster* (Figure 5). However, the values of R for the lowest mean GC content bins are much higher than the values of κ described above. For the lowest mean bins, R is equal to 4.56 and 5.03 in *D. simulans* and *D. melanogaster*, respectively (Figure 5A, B). The values of R are more extreme for the species than the mean bins (Figure 5C, D) – for the lowest species GC content bins, $R = 7.51$ in *D. simulans* and $R = 8.19$ in *D. melanogaster*. The absolute rates $r_{S>W}$ and $r_{W>S}$ are plotted in Supplementary Figure S2. As described above, there are reasons to think that the results for mean bins are more reliable than the species-specific bins, which are likely to be biased towards inferring $S>W$ substitutions. Nevertheless, the fact that $R \gg \kappa$ for the mean bins with low GC content requires an explanation.

One possibility is that there has been selection in these bins in favour of AT rather than GC along both lineages, consistent with the significantly negative value of γ for the lowest GC

content species bin in *D. melanogaster* (Figure 3 D). This would also explain why both the present and reconstructed ancestral GC contents of the low GC bins are much lower than the equilibrium GC content predicted on the basis of neutral evolution and the estimated κ value of approximately 3 (Figure 4); this is equal to $1/[1 + \kappa \exp(-\gamma)]$, which takes a value of 0.25 with $\gamma = 0$ (Li 1987; Bulmer 1991; Charlesworth and Charlesworth 2010, p.275). Table S1 shows that, for the lowest mean bin, the GC content for *D. simulans* is 0.195 and the GC content for *D. melanogaster* is 0.192. The GC contents of the ancestral sequence reconstructed by PAML are 0.197 for the lowest *D. simulans* and *D. melanogaster* mean bins. For neutral equilibrium, such a low GC content requires $\kappa = 4.13$ under neutrality, but a smaller value if there were some selection against GC, e.g. with $\kappa = 2.8$ (the species value for *D. simulans*), an equilibrium GC content of 0.20 requires $\gamma = -0.35$. From Equation 2 of Jackson et al. (2017), this implies an R value of 3.97, which is close to the value estimated for the *D. simulans* lineage. Alternatively, there could have been a shift downwards from an ancestral κ value of approximately 4 in both lineages, which would not be detected if it occurred prior to the period of time for which the polymorphism data are relevant (these only provide an estimate of the value of κ over the past $4N_e$ generations).

Another explanation is that the discrepancy between R and κ for the low GC bins is an artefact due to alignment errors in AT-rich regions, which might contain more repeat sequences (Bachtrog et al. 1999) and thus be harder to align. The *D. melanogaster* reference sequence is moderately more AT-rich in regions that we soft-masked for repeat content compared to the remainder of the genome (GC content for the masked regions = 39%, GC content for the remainder of the genome = 43%). However, the lowest GC content bins do not exhibit lower alignment scores according to the whole genome alignment MULTI-Z output, which suggests that there are no specific alignment problems for these regions

(Supplementary Figure S3). As noted in the methods, we excluded sites masked as repetitive in either species from all of our analyses.

No effect of polymorphism on patterns of divergence

Because the nucleotide diversity within these species represents an appreciable fraction of divergence between them (Table 1), we repeated our analyses of divergence after removing sites that were polymorphic in either the MD sample or the ZI sample, to approximate a dataset consisting only of fixed differences between *D. simulans* and *D. melanogaster*. Removing polymorphic sites had the effect of considerably reducing the substitution rates for both species, more so for *D. simulans* (compare Supplementary Figures S2 and S5 A, B). It had no effect on the patterns of *R* or the substitution count ratio (Supplementary Figure S5).

Discussion

Understanding whether genome evolution involves gBGC or a selection pressure acting on the GC content of putatively functionless sequences is important for two reasons. First, it is needed for a complete understanding of the processes affecting the genetic composition of natural populations. Second, we expect it to affect sites that are often used as comparators for detecting other evolutionary processes, such as selection on functionally important sites, changes in population size and mutation. To date, the evidence for a GC-favouring force in *Drosophila* has been ambiguous. It has been claimed to be acting on the X chromosome of *D. simulans* (Haddrill and Charlesworth 2008) and *D. americana* (de Procé et al. 2012), and on both X chromosomes and autosomes in *D. simulans* and *D. melanogaster* (Jackson et al.

2017), while several other studies have failed to find support for it (Clemente and Vogl 2012; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014). We have extended our previous work on this topic (Jackson et al. 2017) by focussing exclusively on short intronic sites, and by using a larger polymorphism sample in *D. melanogaster* than before, together with more complete annotation of the *D. simulans* genome. This allowed us to take the intersection of sites annotated as short introns in both species, which in turn allowed a direct comparison of the processes acting at homologous sites.

Overall, the analyses presented above suggest the existence of a GC-favouring force in both *D. simulans* and *D. melanogaster*, whose strength is positively related to the GC content of an intron, and which is on average stronger in *D. simulans*. This makes sense in the context of GC-biased gene conversion (gBGC), which is a recombination-association process whose evolutionarily effective strength is proportional to the product of the rate of change of allele frequency by gene conversion and the effective population size (N_e) (Nagylaki 1983; Charlesworth and Charlesworth 2010, p.529). On the basis of pairwise diversity at SI sites, N_e is substantially higher for the *D. simulans* population compared to the *D. melanogaster* population, assuming that mutation rates are similar for the two species (Table 1 and Figure S6), so that the evolutionarily effective strength of any deterministic force over the recent past should be larger in *D. simulans*, other things being equal. There are only weak relationships between SI site diversity and GC content (Figure S6), with an observed ratio of π for the highest versus the lowest species bins of 0.79 for *D. simulans* and 0.84 for *D. melanogaster*. If mutation-selection-drift equilibrium is assumed, and the estimates of κ and γ for these species bins are inserted into Equation 15 of McVean and Charlesworth (1999), the predicted ratios are 0.93 and 1.47 for *D. simulans* and *D. melanogaster*, respectively. The agreement between the observed and predicted values is reasonably good for *D. simulans*, but

there is a large discrepancy for *D. melanogaster*, possibly reflecting a larger departure from base composition equilibrium than in the case of *D. simulans*.

Nevertheless, the presence of a force favouring GC is suggested both by the analyses of polymorphism data, using estimates of the derived allele frequencies of different sorts of mutation (Figure 2) and the site frequency spectrum based estimate of γ (Figure 3), as well as by the analyses of substitution rates (Figure 5). There is also evidence for a strong mutational bias in favour of GC>AT mutations, consistent with direct evidence from mutation accumulation data (Assaf et al. 2017). This bias is larger in *D. melanogaster* than *D. simulans*, and the evidence that base composition is close to statistical equilibrium in the latter but not the former suggests that there may have been a shift towards a stronger mutational bias in the *D. melanogaster* lineage, as has previously been suggested on the basis of somewhat weaker evidence (Kern and Begun 2005; Zeng and Charlesworth 2010; Clemente and Vogl 2012; Jackson et al. 2017). The mechanistic basis and evolutionary significance of such a shift are both unclear.

Our conclusions about the existence of a force favouring GC over AT, and its dependence on the GC content of introns appear to be robust to the different ways of binning SI sites. However, the results also suggest that the conclusions about some of the details of the evolution of SI sites in these two species are sensitive to the methods used to aggregate them, which may be subject to various sources of bias. In particular, binning according to the species-specific GC contents of SIs is likely to produce a bias in favour of W>S substitutions, as discussed in the section on testing for fixation bias, so that the results on substitution patterns from bins based on mean GC content are probably more reliable than those from the species bins.

591

592 In contrast, there has been some divergence in GC content at a subset of the SI sites that are
593 shared between *D. simulans* and *D. melanogaster*, with *D. simulans* SIs having slightly
594 higher GC contents than the same introns in *D. melanogaster* (see the rows labelled ‘Mean’
595 in Table S1). Binning short introns by the mean GC content of homologous sites potentially
596 has the effect of masking some of these differences, because it aggregates sites which are
597 subject to different evolutionary pressures. Use of the species-specific binning method is thus
598 likely to provide a more accurate representation of the sequence context than the use of
599 means, which is relevant to estimates of the strength of gBGC or selection from
600 polymorphism data, although these are qualitatively consistent with the patterns from the
601 mean bins.

602

603 We explored these differences further by additionally binning sites by the *difference* in GC
604 content between *D. simulans* and *D. melanogaster* (GC_{diff}), and observed even more extreme
605 patterns than for the species bins (Supplementary Figure S4), with evidence from the
606 polymorphism data for selection against GC in *D. simulans* in bins for which GC_{diff} is
607 smallest (and negative), and against GC in *D. melanogaster* in bins for which this difference
608 is the largest. Both R and γ change with increasing GC_{diff} in opposite directions in the two
609 species (Figure S4). The relationship between R and GC_{diff} is probably partly due to the
610 sampling bias described in the section on testing for fixation bias: selecting introns on the
611 basis of GC_{diff} causes a bias towards an excess of W>S changes as opposed to S>W changes
612 in the *D. simulans* lineage for introns with a low GC content in *D. simulans*, and vice-versa
613 for *D. melanogaster*.

614

615 The complementary relation observed for γ is more difficult to explain, since the same sets
616 of introns in each bin are being used for the polymorphism analyses in the two species, yet
617 low GC_{diff} introns have low γ values in *D. simulans* and high γ values in *D. melanogaster*.
618 This finding implies that shifts in the base composition of a considerable fraction of introns
619 must have occurred along each lineage as a result of changes in γ values in opposite
620 directions in the two lineages. This possibly even includes shifts in the direction of the
621 directional force or forces affecting the frequencies of GC versus AT basepairs, given the
622 evidence for negative γ in low GC_{diff} introns in *D. simulans* and in high GC_{diff} introns in *D.*
623 *melanogaster* (Figure S4). Such changes would also contribute to the pattern for R described
624 above; Equation 2 of Jackson *et al.* (2017) shows that, with $\kappa = 3$, a γ of -1 results in $R =$
625 8.89. It is, however, surprising that they have been sufficiently large to produce shifts in GC
626 content of approximately 0.10 over the relatively short evolutionary timescale involved.
627
628 The biological mechanisms underlying the GC-favouring force, if it is real, are unclear.
629 Direct experimental evidence for a GC-bias in transmission of alleles to the products of
630 meiosis due to the repair machinery associated with recombination is limited to budding
631 yeast, where the segregation distortion in favour of GC is modest (Mancera *et al.* 2008; but
632 see Liu *et al.* 2018) and to mammals, where the distortion is strong at recombination hotspots
633 (Webb *et al.* 2008; Duret and Galtier 2009; Arbeithuber *et al.* 2015). Indeed, Liu *et al.* (2018)
634 failed to detect significantly GC-biased segregation in yeast, *Neurospora*, *Chlamydomonas*
635 and *Arabidopsis*, although population genetics evidence has suggested its existence in yeast
636 (Harrison and Charlesworth 2011) and *Arabidopsis* (Hämälä and Tiffin 2020). To our
637 knowledge, there is no direct experimental evidence for gBGC, or any other form of biased
638 transmission in *Drosophila*. In mammals, it has been hypothesised that the strength of gBGC
639 is an adaptation to counter the high rate of mutation of methylated cytosines (Brown and

Jiricny 1987; Duret and Galtier 2009). *Drosophila* has far lower levels of cytosine methylation compared to mammals (Gowher et al. 2000; Capuano et al. 2014), and the mismatch repair machinery of *Drosophila* differs from mammals and other eukaryotes in important ways (Sekelsky 2017). Consequently, it is unclear *a priori* what level of expectation there is for a GC or an AT bias in transmission of alleles in *Drosophila*. Direct observation of the progenitors and products of meiosis in *Drosophila* would be useful for testing the patterns reported here. Given that there is some doubt about the accuracy of ancestral state inference when the standard assumption of neutral evolution is applied, even using the up-to-date methods employed here, we suggest that work extending models of base composition evolution to incorporate weak directional forces (such as gBGC) would also be worthwhile (e.g., Borges et al. 2019).

Acknowledgements

The authors thank Kai Zeng for helpful comments on the manuscript, and Henry Barton for providing the multiple alignment between *Drosophila* species.

References

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

663 Akashi H, Ko W-Y, Piao S, John A, Goel P, Lin C-F, Vitins AP. 2006. Molecular evolution
664 in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the
665 timescale of molecular divergence. *Genetics* 172:1711-1726.

666 Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149-
667 1152.

668 Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated
669 with mutation and biased gene conversion at recombination hotspots. *Proceedings of the*
670 *National Academy of Sciences* 112:2109-2114.

671 Assaf ZJ, Tilk S, Park J, Siegal ML, Petrov DA. 2017. Deep sequencing of natural and
672 experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new
673 mutations. *Genome Research* 27:1988-2000.

674 Bachtrog D, Weiss S, Zangerl B, Brem G, Schlötterer C. 1999. Distribution of dinucleotide
675 microsatellites in the *Drosophila melanogaster* genome. *Molecular Biology and Evolution*
676 16:602-610.

677 Barton HJ, Zeng K. 2018. New methods for inferring the distribution of fitness effects for
678 indels and snps. *Molecular Biology and Evolution* 35:1536-1546.

679 Becher H, Jackson BC, Charlesworth B. 2020. Patterns of genetic variability in genomic
680 regions with low rates of recombination. *Curr Biol* 30:94-100 e103.

681 Begun DJ. 2001. The frequency distribution of nucleotide variation in *Drosophila simulans*.
682 *Molecular Biology and Evolution* 18:1343-1352.

683 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R,
684 Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences
685 with the threaded blockset aligner. *Genome Research* 14:708-715.

686 Borges R, Szöllösi GJ, Kosiol C. 2019. Quantifying gc-biased gene conversion in great ape
687 genomes using polymorphism-aware models. *Genetics* 212:1321-1336.

688 Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE,
689 Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR. 2008. Assessing the evolutionary impact of
690 amino acid mutations in the human genome. *PLoS Genet* 4:e1000083.

691 Brown TC, Jiricny J. 1987. A specific mismatch repair event protects mammalian cells from
692 loss of 5-methylcytosine. *Cell* 50:945-950.

693 Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*
694 129:897-907.

695 Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining
696 regions of the *Drosophila melanogaster* genome. *Genome Biology and Evolution* 4:278-288.

697 Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2013. Codon usage bias and
698 effective population sizes on the X chromosome versus the autosomes in *Drosophila*
699 *melanogaster*. *Molecular Biology and Evolution* 30:811-823.

700 Capuano F, Mülleder M, Kok R, Blom HJ, Ralser M. 2014. Cytosine DNA methylation is
701 found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*,
702 *Schizosaccharomyces pombe*, and other yeast species. *Analytical chemistry* 86:3697-3702.

703 Charlesworth B, Charlesworth D. 2010. *Elements of Evolutionary Genetics*. Greenwood
704 Village: Roberts and Company Publishers.

705 Clemente F, Vogl C. 2012. Unconstrained evolution in short introns? - an analysis of
706 genome-wide polymorphism and divergence data from *Drosophila*. Journal of Evolutionary
707 Biology 25:1975-1990.

708 Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in
709 *Drosophila melanogaster*. PLOS Genetics 8:e1002905-e1002905.

710 de Procé SM, Zeng K, Betancourt AJ, Charlesworth B. 2012. Selection on codon usage and
711 base composition in *Drosophila americana*. Biology letters 8:82-85.

712 Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic
713 landscapes. Annual Review of Genomics and Human Genetics 10:285-311.

714 Efron B. 1979. Bootstrap methods: Another look at the jackknife. In: Institute of
715 Mathematical Statistics. p. 1-26.

716 Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in
717 the presence of slightly deleterious mutations and population size change. Molecular Biology
718 and Evolution 26:2097-2108.

719 Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new
720 deleterious amino acid mutations in humans. Genetics 173:891-900.

721 Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL,
722 Excoffier L. 2007. Statistical evaluation of alternative models of human evolution.
723 Proceedings of the National Academy of Sciences of the United States of America
724 104:17614-17619.

725 Fay JC, Wyckoff GJ, Wu C-I. 2001. Positive and negative selection on the human genome.
726 Genetics 158:1227-1234.

727 Galtier N. 2016. Adaptive protein evolution in animals and the effective population size
728 hypothesis. PLOS Genetics 12.

729 Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in north
730 american *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet
731 11:e1005004-e1005004.

732 Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of gc-
733 biased gene conversion in the human genome. Genome Research 25:1215-1228.

734 Gowher H, Leismann O, Jeltsch A. 2000. DNA of *Drosophila melanogaster* contains 5-
735 methylcytosine. The EMBO journal 19:6918-6923.

736 Green MR. 1986. Pre-mrna splicing. Annual Review of Genetics 20:671-708.

737 Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding
738 DNA in *Drosophila simulans*. Molecular Biology and Evolution 25:1825-1834.

739 Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition
740 of non-coding sequences in *Drosophila*. Biology letters 4:438-441.

741 Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of
742 nucleotide variability and the demographic and selection history of *Drosophila melanogaster*
743 populations. Genome Research 15:790-799.

744 Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome
745 revealed by a genome-wide interspecies comparison. Genome Research 16:875-884.

746 Härmälä T, Tiffin P. 2020. Biased gene conversion constrains adaptation in *Arabidopsis*
747 *thaliana*. Genetics.

748 Harris RS. 2007. Improved pairwise alignment of genomic DNA. [PhD thesis]: The
749 Pennsylvania State University.

750 Harrison RJ, Charlesworth B. 2011. Biased gene conversion affects patterns of codon usage
751 and amino acid usage in the *Saccharomyces sensu stricto* group of yeasts. *Molecular Biology*
752 *and Evolution* 28:117-129.

753 Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation
754 rates may cause spurious signatures of a fixation bias favoring higher gc-content in humans.
755 *Molecular Biology and Evolution* 24:2196-2202.

756 Jackson BC, Campos JL, Haddrill PR, Charlesworth B, Zeng K. 2017. Variation in the
757 intensity of selection on codon bias over time causes contrasting patterns of base composition
758 evolution in *Drosophila*. *Genome Biology and Evolution* 9.

759 Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-
760 Stabelli O, Kankare M, Bogaerts-Márquez M. 2020. Genomic analysis of European
761 *Drosophila melanogaster* populations reveals longitudinal structure, continent-wide
762 selection, and previously unknown DNA viruses. *Molecular Biology and Evolution* 37:2661-
763 2678.

764 Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of
765 deleterious mutations and population demography based on nucleotide polymorphism
766 frequencies. *Genetics* 177.

767 Keightley PD, Jackson BC. 2018. Inferring the probability of the derived vs. the ancestral
768 allelic state at a polymorphic site. *Genetics* 209:897-906.

769 Kennedy CF, Berget SM. 1997. Pyrimidine tracts between the 5'splice site and branch point
770 facilitate splicing and recognition of a small *Drosophila* intron. Molecular and Cellular
771 Biology 17:2774-2780.

772 Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron:
773 Duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of
774 the National Academy of Sciences of the United States of America 100:11484-11489.

775 Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding
776 sequences of *Drosophila melanogaster* and *D. simulans*: Evidence for nonequilibrium
777 processes. Molecular Biology and Evolution 22:51-62.

778 Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the distribution of selection
779 coefficients for new nonsynonymous mutations using large samples. Genetics 206:345-361.

780 Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217:624-626.

781 Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley
782 CH, Pool JE. 2015. The drosophila genome nexus: A population genomic resource of 623
783 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population.
784 Genetics 199:1229-1241.

785 Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at
786 synonymous sites in *D. melanogaster*. PLOS Genetics 9:e1003527-e1003527.

787 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
788 transform. Bioinformatics (Oxford, England) 25:1754-1760.

789 Li W-H. 1987. Models of nearly neutral mutations with particular implications for
790 nonrandom usage of synonymous codons. Journal of Molecular Evolution 24:337-345.

791 Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. 2018. Tetrad
792 analysis in plants and fungi finds large differences in gene conversion rates but no GC bias.
793 Nature Ecology & Evolution 2:164-173.

794 Machado HE, Lawrie DS, Petrov DA. 2020. Pervasive strong selection at the level of codon
795 usage bias in *Drosophila melanogaster*. Genetics 214:511-528.

796 Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping
797 of meiotic crossovers and non-crossovers in yeast. Nature 454:479-485.

798 Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of ancestral sequence reconstruction
799 methods to infer nonstationary patterns of nucleotide substitution. Genetics 200:873-890.

800 McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *adh* locus in *Drosophila*.
801 Nature 351:652-654.

802 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
803 Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A mapreduce
804 framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-
805 1303.

806 McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of
807 synonymous codon usage: Patterns and predictions. Genetical Research 74:145-158.

808 Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test.
809 Proceedings of the National Academy of Sciences of the United States of America 110:8615-
810 8620.

811 Mount SM, Burks C, Herts G, Stormo GD, White O, Fields C. 1992. Splicing signals in
812 *Drosophila*: Intron size, information content, and consensus sequences. Nucleic Acids
813 Research 20:4255-4262.

814 Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proceedings of the
815 National Academy of Sciences 80:6278-6281.

816 Obbard DJ, MacLennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012.
817 Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. Molecular
818 Biology and Evolution 29:3459-3473.

819 Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility
820 of short intron sequences as a reference for the detection of positive and negative selection in
821 *Drosophila*. Molecular Biology and Evolution 27:1226-1234.

822 Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchon P,
823 Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population genomics of sub-saharan *Drosophila*
824 *melanogaster*: African diversity and non-African admixture. PLOS Genetics 8:e1003080-
825 e1003080.

826 Quinlan AR, Hall IM. 2010. Bedtools: A flexible suite of utilities for comparing genomic
827 features. Bioinformatics 26:841-842.

828 Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence
829 for GC-biased gene conversion in *Drosophila melanogaster*. Molecular Biology and
830 Evolution 31:425-433.

831 Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of
832 standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*.
833 Molecular Biology and Evolution 31:1750-1766.

834 Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring
835 the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427-
836 1437.

837 Sekelsky J. 2017. DNA repair in *Drosophila*: Mutagens, models, and missing genes. Genetics
838 205:471-490.

839 Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature
840 415:1022-1024.

841 Takano-Shimizu T. 2001. Local changes in GC/AT substitution biases and in crossover
842 frequencies on *Drosophila* chromosomes. Molecular Biology and Evolution 18:606-619.

843 Tange O. 2011. Gnu parallel-the command-line power tool. The USENIX Magazine 36:42-
844 47.

845 Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects
846 and proportion of adaptive substitutions from polymorphism data. Genetics 207:1103-1119.

847 R Core Team. 2018. R: A language and environment for statistical computing. Vienna,
848 Austria: R Foundation for Statistical Computing.

849 Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*.
850 BMC Evolutionary Biology 7:226-226.

851 Webb AJ, Berg IL, Jeffreys A. 2008. Sperm cross-over activity in regions of the human
852 genome showing extreme breakdown of marker association. *Proceedings of the National*
853 *Academy of Sciences* 105:10471-10476.

854 Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*.
855 *Genetics* 173:821-837.

856 Wright S. 1938. The distribution of gene frequencies under irreversible mutation.
857 *Proceedings of the National Academy of Sciences of the United States of America* 24:253.

858 Yang Z. 2007. Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology*
859 *and Evolution* 24:1586-1591.

860 Zeng K, Charlesworth B. 2010. Studying patterns of recent evolution at synonymous sites
861 and intronic sites in *Drosophila melanogaster*. *Journal of Molecular Evolution* 70:116-128.

862 Zeng K, Jackson BC, Barton HJ. 2019. Methods for estimating demography and detecting
863 between-locus differences in the effective population size and mutation rate. *Molecular*
864 *Biology and Evolution* 36:423-433.

865

866

867

868 **Tables**

869

870 Table 1. Polymorphism statistics for SI sites (means with their 95% CIs)

871

Population	Nucleotide diversity (π)	Watterson's θ	Tajima's D	Proportion of singletons
MD21	0.0371 (0.0365, 0.0377)	0.0513 (0.0507, 0.0521)	-1.16 (-1.133, -1.183)	0.513 (0.506, 0.519)
ZI69	0.0164 (0.0159, 0.0168)	0.0195 (0.0191, 0.0199)	-0.566 (-0.517, -0.616)	0.308 (0.301, 0.317)

872

873

874

Figures

Figure 1. The ratios of substitution counts, plotted against the GC content of SI bins. A substitution count ratio of $N_{W>S}/N_{S>W} = 1$ implies equilibrium base composition. Ratios were calculated for the *D. simulans* lineage (panels A; C) and the *D. melanogaster* lineage (B; D). Sites were binned by the mean GC content of homologous introns in *D. melanogaster* and *D. simulans* (A; B) or by the species-specific GC content (C; D). Error bars represent 95% CIs from 1000 bootstraps of the data in each bin.

885

886

Figure 2. Derived allele frequencies (DAF) for different types of mutations, plotted against the GC content of SI bins. DAF was calculated for the *D. simulans* MD population (panels A; C) and the *D. melanogaster* ZI population (B; D). Sites were binned by the average GC content of homologous introns in *D. melanogaster* and *D. simulans* (A; B) or by species-specific GC content (C; D). Error bars represent 95% CIs from 1000 bootstraps of the data in each bin.

Figure 3. Estimates of the strength of selection in favour of GC alleles ($\gamma = 4N_e s$), plotted against the GC content of SI bins. Open symbols indicate values which are not significantly different from zero. Circular symbols denote models incorporating parameters that correct for polarisation error. Triangular symbols denote models not incorporating parameters that correct for polarisation error. γ was calculated for *D. simulans* (panels A; C) and *D. melanogaster* (panels B; D). Sites were binned by the mean GC content of homologous introns in *D. melanogaster* and *D. simulans* (panels A; B) or by the species-specific GC content (panels C; D).

Figure 4. The estimates of the mutational bias parameter, κ , calculated using the method of Glémin *et al.* (2015), plotted against the GC content of SI bins. κ was estimated from the model M1 (with $\gamma \neq 0$ and polarisation errors not corrected for), for both *D. simulans* (panels A; C) and the *D. melanogaster* (B; D). Sites were binned by the mean GC content of homologous introns in *D. melanogaster* and *D. simulans* (A; B) or by the species-specific GC content (C; D).

Figure 5. The ratio of substitution rates, $R = r_{S \rightarrow W} / r_{W \rightarrow S}$ plotted against the GC content of SI bins. R was calculated for the *D. simulans* lineage (panels A; C) and the *D. melanogaster* lineage (B; D). Sites were binned by the average GC content of homologous introns in *D. melanogaster* and *D. simulans* (A; B) or by species-specific GC content (C; D). Error bars represent 95% CIs from 1000 bootstraps of the data in each bin.

928

929