Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in *Drosophila*

Benjamin C Jackson[1], José L Campos[2], Penelope R Haddrill[3], Brian Charlesworth[2] and Kai Zeng[1*]

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, United Kingdom

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

[3] Centre for Forensic Science, Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, G1 1XW, United Kingdom

[*]Author for correspondence: Kai Zeng, Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom, +44 (0) 114 222 4708, k.zeng@sheffield.ac.uk

Abstract

Four-fold degenerate coding sites form a major component of the genome, and are often used to make inferences about selection and demography, so that understanding their evolution is important. Despite previous efforts, many questions regarding the causes of base composition changes at these sites in *Drosophila* remain unanswered. To shed further light on this issue, we obtained a new whole-genome polymorphism dataset from *D. simulans*. We analysed samples from the putatively ancestral range of *D. simulans*, as well as an existing polymorphism dataset from an African population of *D. melanogaster*. By using *D. yakuba* as an outgroup, we found clear evidence for selection on 4-fold sites along both lineages over a substantial period, with the intensity of selection increasing with GC content. Based on an explicit model of base composition evolution, we suggest that the observed AT-biased substitution pattern in both lineages is probably due to an ancestral reduction in selection intensity, and is unlikely to be the result of an increase in mutational bias towards AT alone. By using two polymorphism-based methods for estimating selection coefficients over different timescales, we show that the selection intensity on codon usage has been rather stable in *D. simulans* in the recent past, but the long-term estimates in *D. melanogaster* are much higher than the short-term ones, indicating a continuing decline in selection intensity, to such an extent that the short-term estimates suggest that selection is only active in the most GC-rich parts of the genome. Finally, we provide evidence for complex evolutionary patterns in the putatively neutral short introns, which cannot be explained by the standard GC-biased gene conversion model. These results reveal a dynamic picture of base composition evolution.

Key words

Codon usage bias, non-equilibrium behaviour, selection, short introns, *Drosophila*

Introduction

Here, we investigate the forces that affect evolution at 4-fold degenerate coding sites in *Drosophila simulans* and *D. melanogaster*. These sites represent a substantial part of the genome, and are often used as references against which selection at other sites, for example non-synonymous sites, is tested (McDonald & Kreitman 1991; Rand & Kann 1996; Parsch et al. 2010; Stoletzki & Eyre-Walker 2011). Quantifying the forces that affect their evolution is necessary both for a general understanding of genome evolution and for making robust inferences about the influences of demographic factors and selection elsewhere in the genome (Matsumoto et al. 2016).

Codon usage bias (CUB) is a key feature of 4-fold sites, since it involves the disproportionate use of certain codons among the set of codons that code for a given amino acid. There is evidence for CUB in a wide range of organisms, including both prokaryotes and eukaryotes (Drummond & Wilke 2008; Hershberg & Petrov 2008). The most common explanation for CUB is that this maximises translational efficiency and/or accuracy (Hershberg & Petrov 2008). Avoidance of the toxicity of misfolded proteins generated by translational errors has also been proposed as an explanation of CUB (Drummond & Wilke 2008). Recent work has also suggested the possibility that stabilizing, as opposed to directional, selection maintains the frequencies of synonymous codons, because CUB has been found to be unrelated to recombination rate in *D. pseudoobscura*, in line with theoretical predictions about the action of stabilizing selection (Charlesworth 2013; Fuller et al. 2014; Kliman 2014).

In most species of *Drosophila* for which data are available, including *D. melanogaster* and *D. simulans*, all the preferred codons are GC-ending (Vicario et al. 2007; Zeng 2010). Selection

for preferred codons thus acts to increase the GC content of third position sites in coding sequences, and GC-ending and AT-ending codons have been often used as proxies for preferred and unpreferred codons, respectively. As in other species, evidence for selection for preferred codons in *D. melanogaster* comes from the fact that the level of codon bias is related to expression level (e.g., Duret & Mouchiroud 1999; Hey & Kliman 2002; Campos et al. 2013). There is also a negative relationship between the level of CUB and synonymous site divergence in the *Drosophila melanogaster* subgroup, consistent with selection for preferred codons (Shields et al. 1988; Powell & Moriyama 1997; Dunn et al. 2001; Bierne & Eyre-Walker 2006).

However, analyses based on between-species sequence divergence have consistently revealed an excess of substitutions towards AT-ending codons in the *D. melanogaster* lineage (Akashi 1995, 1996; McVean & Vieira 2001; Poh et al. 2012). Two hypotheses have been proposed for this observation. These are, firstly, that *D. melanogaster* has undergone a reduction in the population-scaled strength of selection for preferred codons, $4N_e s$, where N_e is the effective population size and s is the selection coefficient favouring preferred codons in heterozygotes for the preferred allele. This reduction in selection could be caused either by a reduction in N_e (Akashi 1996), or a reduction in s, perhaps due to changed ecological conditions (Clemente & Vogl 2012a, 2012b). The second explanation is that *D. melanogaster* has undergone a shift in mutational bias towards AT alleles (Takano-Shimizu 2001; Kern & Begun 2005; Zeng & Charlesworth 2010a; Clemente & Vogl 2012b). It has also been argued that both factors must be invoked to explain patterns of variation and evolution in the *D. melanogaster* lineage (Nielsen et al. 2007; Clemente & Vogl 2012a, 2012b).

Several attempts to detect selection on codon bias in *D. melanogaster* have come to conflicting conclusions. For instance, some polymorphism-based studies managed to detect evidence for selection favouring GC-ending codons (Zeng & Charlesworth 2009; Campos et al. 2013), although the intensity of selection may be weak relative to other *Drosophila* species (Kliman 1999; Andolfatto et al. 2011). However, other studies did not find support for such ongoing selection (Clemente & Vogl 2012a; Vogl & Clemente 2012; Poh et al. 2012). Thus, there is a pressing need to gain a better understanding of the dynamics of selection on codon bias and understand the sources of these conflicting results.

Much less is known about *D. simulans*. Early studies based on a small number of loci suggest that this species may be at base composition equilibrium, with the number of substitutions from AT-ending codons to GC-ending codons not statistically different from that in the opposite direction (e.g., Akashi 1995, 1996; Kern & Begun 2005; Akashi et al. 2006; Haddrill & Charlesworth 2008). However, more recent analyses have revealed AT-biased substitution patterns (Begun et al. 2007; Poh et al. 2012), suggesting a possible reduction in selection intensity in this lineage, although the reduction may be less severe compared to that in *D. melanogaster* (McVean & Vieira 2001). In contrast to the situation in *D. melanogaster*, the few polymorphism-based studies in *D. simulans* generally point to evidence for selection for preferred codons (Akashi 1997, 1999; Kliman 1999; Andolfatto et al. 2011). It is therefore unclear whether/how selection intensity has changed over time in *D. simulans*, and how the dynamics of base composition evolution differ from those in *D. melanogaster*.

Irrespective of the reason(s) for the AT-biased substitution pattern in these two *Drosophila* lineages, these findings present a problem for ancestral state reconstruction, a process that is necessary for inferring substitution patterns along a lineage of interest and for polarising

segregating sites into ancestral and derived variants to understand their more recent evolution. Use of maximum parsimony methods or maximum likelihood models that assume equilibrium base composition under such circumstances can lead to erroneous inferences, although these two methods were used in many previous analyses of various *Drosophila* species (Akashi et al. 2007; Matsumoto et al. 2015). Departures from base composition equilibrium may also lead to complex polymorphism patterns (Zeng & Charlesworth 2009). Both of these sources of difficulties may contribute to the mixed evidence for the nature of the forces acting on synonymous sites in *Drosophila* (Zeng & Charlesworth 2010a; Clemente & Vogl 2012a).

A factor that may confound the study of CUB is GC-biased gene conversion (gBGC), which is a recombination-associated process, and acts to increase GC content at sites where recombination occurs (Duret & Galtier 2009). Most studies have found little or no evidence for gBGC in *D. melanogaster* (Clemente & Vogl 2012b; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014), although there is some evidence either for the action of selection for GC basepairs or gBGC on the evolution of non-coding sequences in *D. simulans* (Haddrill & Charlesworth 2008). In order to control for gBGC, we have analysed data on the 8-30bp region of short introns (SIs), which are widely considered to be evolving near-neutrally in *Drosophila* (Halligan & Keightley 2006; Parsch et al. 2010; Clemente & Vogl 2012b).

To address the questions raised above, we need to look at both divergence and polymorphism data from both species; the analyses should explicitly take into account departures from equilibrium, so that signals of selection can be detected without biases. To this end, we have obtained new whole-genome data from *D. simulans*, and used an existing high-quality dataset for *D. melanogaster*. Using the reference genome of *D. yakuba* as an outgroup, we used state-

of-the-art methods to reconstruct ancestral states. In addition, we employed methods that can infer selection intensity on different timescales, along the *D. melanogaster* and *D. simulans* lineages, with the aim of shedding further light on the evolutionary dynamics of genome composition in these two species.

Material and Methods

Sequence data preparation

We first describe the sequencing of 22 new *D. simulans* isofemale lines, 11 of which were collected by William Ballard in 2002 from Madagascar (MD lines – MD03, MD146, MD197, MD201, MD224, MD225, MD235, MD238, MD243, MD255, MD72); the other 11 were collected by Peter Andolfatto in 2006 from Kenya (NS lines – NS11, NS111, NS116, NS19, NS37, NS49, NS63, NS64, NS89, NS95, NS96). We produced homozygous lines by full-sib inbreeding in the Charlesworth lab for nine generations; however, six lines (NS11, NS63, NS116, MD224, MD243, MD255) were lost early in the process of inbreeding. For these lines, we sequenced the initial stocks that we had received from the Andolfatto lab. Genomic DNA was prepared for each isofemale line by pooling twenty-five females, snap freezing them in liquid nitrogen, extracting DNA using a standard phenol-chloroform extraction protocol with ethanol, and ammonium acetate precipitation. These flies were sequenced by the Beijing Genomics Institute (BGI; http://bgi-international.com/). A 500bp short-insert library was constructed for each sample, and the final data provided consisted of 90bp paired-end Illumina sequencing (pipeline version 1.5), with an average coverage of 64X. We double-checked the quality of the filtered reads for each allele with FastQC (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and no further trimming was necessary. The raw reads have been deposited in the European Nucleotide Archive, study accession number: PRJEB7673.

We obtained sequence data for 20 further *D. simulans* isofemale lines from Rogers et al. (2014). These lines were from the same sampling localities in Kenya (10 lines: NS05, NS113, NS137, NS33, NS39, NS40, NS50, NS67, NS78, NS79) and Madagascar (10 lines: MD06,

MD105, MD106, MD15, MD199, MD221, MD233, MD251, MD63, MD73) as above. Each line was sequenced on between 2-3 lanes of paired-end Illumina sequencing at the UC Irvine High Throughput Genomics centre (http://ghtf.biochem.uci.edu/) per line. Further information about these lines and their sequencing is available in Rogers et al. (2014). After examining FastQC files for these 20 lines, we trimmed two lines with apparently lower quality scores (MD233 and MD15) using the trim-fastq.pl script from Popoolation 1.2.2 (Kofler et al. 2011) with the (minimum average per base quality score) --quality-threshold flag set to 20.

Downstream of sequencing, we combined both datasets and used a BWA/SAMtools/GATK pipeline, previously described in Campos et al. (2014) and Jackson et al. (2015), to generate genotype calls. Briefly, we aligned and mapped reads for each *D. simulans* line to the second generation assembly of the *D. simulans* reference sequence (Hu et al. 2013) using BWA 0.7.10 (Li & Durbin 2009). We used SAMtools 1.1 (Li et al. 2009) to filter alignments with a mapping quality < 20, and to sort and index the resulting alignments. To combine reads from one sample across multiple lanes, we used Picard tools 1.119 (http://broadinstitute.github.io/picard/) to edit BAM file headers and SAMtools 1.1 to merge, resort and index BAM files *per* sample. We then used Picard tools 1.119 to fix mate information, sort the resulting BAM files and mark duplicates. We performed local realignment using the RealignerTargetCreator and IndelRealigner tools of GATK 3.3 (https://www.broadinstitute.org/gatk/).

For SNP calling, we used the UnifiedGenotyper for diploid genomes (parameter: sample_ploidy 2) and generated a multisample VCF file (Danecek et al. 2011). Subsequently, we performed variant quality score recalibration (VQSR) to separate true variation from

machine artefacts (DePristo et al. 2011). We used biallelic and homozygous (for a given individual) SNPs detected at 4-fold sites at a frequency equal to or higher than seven sequenced individuals as the training set. Six SNP call annotations were considered by the VQSR model: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, and MQ, as suggested by GATK (see http://www.broadinstitute.org/gatk/; DePristo et al. 2011). The SNPs were allocated to tranches according to the recalibrated score, so that a given proportion of the true sites were recovered. We retained variants that passed a cutoff of 95%, the variant score limit that recovers 95% of the variants in the true data set. We refer to this dataset as 'filtered'. From the multisample recalibrated VCF file, we made a consensus sequence FASTA file for each individual using a custom Perl script. The variant calls that did not pass the filter were called N (missing data) at the sites in question. We also generated an unfiltered dataset, where we did not implement any form of variant score recalibration. We refer to this dataset as 'unfiltered'. The VCF files and the scripts used to produce them can be downloaded by following the hyperlink provided on http://zeng-lab.group.shef.ac.uk.

Annotation of the *D. simulans* dataset

Using annotations from the *D. simulans* reference (Hu et al. 2013) we extracted coding sequence (CDS) for each gene and made FASTA alignments. We included the *D. simulans* reference sequence as well as the 1:1 FlyBase orthologous genes of *D. melanogaster* (release version 5.33) and *D. yakuba* (release version 1.3). We then performed amino-acid sequence alignments using MAFFT (Katoh et al. 2002). These amino-acid sequence alignments were translated back to nucleotides using custom scripts in PERL to produce in-frame coding sequence alignments that included the 42 *D. simulans* alleles and the *D. melanogaster* and the *D. yakuba* outgroups. We extracted 4-fold (and 0-fold) degenerate sites from CDS alignments which were 4-fold (0-fold) degenerate in all lines, with the condition that there was at most

one segregating site in the codon to which the 4-fold (0-fold) site belonged. We retained the 4-fold (0-fold) sites from an alignment only if there were at least ten 4-fold (0-fold) sites in that alignment in total. For the polymorphism and substitution analyses on 4-fold sites reported in the Results, we carried out the same procedure with the added condition that sites must also be 4-fold degenerate in the three reference sequences.

We also extracted the intron coordinates from the *D. simulans* reference genome sequence. Genomes were masked for any possible exons. For each *D. simulans* intron, we obtained the corresponding orthologous intron of *D. melanogaster* (Hu et al. 2013). For *D. yakuba*, for each orthologous gene, we obtained all its annotated introns and blasted them against the *D. melanogaster* introns (of the same ortholog) with an e-value of less than 10^{-5} and selected the reciprocal best hit (because introns are generally short, the threshold e-value was conservative; see Results). We used RepeatMasker (http://www.repeatmasker.org) to mask repetitive elements in our intron dataset, using the library of repeats for *D. melanogaster* and the default settings. We produced a final alignment of each intronic polymorphic dataset of *D. simulans* with the corresponding *D. melanogaster* and *D. yakuba* orthologs using MAFFT.

We extracted positions 8-30bp of all introns < 66bp long, based on the *D. melanogaster* reference alignment for each intron, as we considered the *D. melanogaster* reference to be the best annotated of the three species. To do this, we scanned the *D. melanogaster* reference sequence for each intronic alignment. We retained the alignment if the *D. melanogaster* reference sequence was less than 66bp long (not including alignment gaps), and then further obtained the coordinates of the 8bp position and the 30bp position in the *D. melanogaster* reference sequence after discarding any gaps introduced by the alignment program. We then cut the whole alignment at these coordinates. These short intronic (SI) sites are thought to be

close to neutrally evolving in *Drosophila*, based on their patterns of polymorphism and substitution (Halligan & Keightley 2006; Parsch et al. 2010; Clemente & Vogl 2012b).

The *D. melanogaster* dataset

Similar analyses were performed using a *D. melanogaster* polymorphism dataset, described in Jackson et al. (2015), which consists of 17 Rwandan *D. melanogaster* samples (RG18N, RG19, RG2, RG22, RG24, RG25, RG28, RG3, RG32N, RG33, RG34, RG36, RG38N, RG4N, RG5, RG7 and RG9) made available by the *Drosophila* Population Genomics Project 2 (http://www.dpgp.org/dpgp2/candidate/).

Quality control of *D. simulans* genotypes

The lines that were inbred successfully for nine generations to produce homozygous samples still retained low levels of residual heterozygosity, which may have been due to a failure to purge our lines of natural variation (Stone 2012), or to SNP calling errors (the latter should be less likely given the high coverage [64x] and our stringent SNP calling regime). We quantified the amount of residual heterozygosity per sample for each of the unfiltered and filtered datasets (Supplementary figure S1). As expected, the filtered dataset exhibited lower levels of residual heterozygosity (ND samples: mean value = 0.0616%, all values < 0.5%; MD samples: mean value = 0.0168%, all values < 0.15%). The six lines that were not subject to the inbreeding procedure (see above) did not have substantially higher levels of residual heterozygosity than the remaining samples, presumably because they were already considerably inbred after being kept as laboratory stocks for several years. For downstream analyses we treated heterozygous sites 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. Because our samples are from partially inbred lines that originated from a mating between at least one wild male and only one wild female, heterozygosity at a site implies that the site is segregating in the wild population. By sampling one allele at random, we attempted to replicate the inbreeding process, which aimed to remove heterozygosity from within the lines.

Pairwise π_S values (synonymous site diversity) for all 42 *D. simulans* lines showed three pairs of samples which deviated substantially from the distribution of pairwise π_S between samples (mean π_S for all samples = 0.030, s.d. = 0.0018). These pairs were MD201—NS116 (π_S = 7.28 x 10^{-5}); NS137—NS37 (π_S = 0.0034) and NS49—NS96 (π_S = 0.0097). A PCA of binary genotypes placed NS116 within the cluster of MD samples, and NS116 exhibited a more MD-like genetic distance to the *D. simulans* reference sequence. These results were based on the filtered dataset, but the unfiltered dataset returned qualitatively identical patterns (data not shown). We therefore excluded NS116 from all downstream analyses based on the likelihood of its representing labelling error. We also excluded NS37 and NS96 as these individuals had the highest levels of residual heterozygosity out of the remaining two pairs of closely related samples (Supplementary figure S1).

To further assess the quality of our datasets, we compared polymorphism and divergence statistics to data previously published in the literature on *D. simulans* (see Results). In particular, we calculated a range of summary statistics per gene: F_{ST} between NS and MD samples; π, Tajima's D, Δ_π, and θ_W within the NS sample, within the MD sample, and for both samples combined. Δ_π for a given gene (Langley et al. 2014) is defined as

$$\Delta_\pi = \frac{\hat{k}}{S} - \frac{1}{\sum_{i=1}^{n-1}(1/i)} \quad (1)$$

where k represents the mean number of pairwise differences among the n alleles in the sample, and S is the number of segregating sites (Langley et al. 2014). We calculated this statistic using a modified version of the tajima.test() function from the pegas package (Paradis 2010) in R. Δ_π is similar to Tajima's D (Tajima 1989), but is normalised by the total amount of diversity. Its advantage over Tajima's D is that it is less dependent on the total diversity for the sample (Langley et al. 2014). We also compared K_A and K_S between the three reference sequences (*D. melanogaster, D. simulans* and *D. yakuba*) in all CDS alignments using the kaks() function from the seqinr package in R, and K_{SI} between the reference sequences in all our short intronic alignments using the dist.dna() function from the pegas package in R, based on the K80 method (Kimura 1980). These analyses are presented in the first section of the Results.

Divergence-based analyses

We used three methods to determine the ancestral state at the *melanogaster-simulans* (*ms*) node, all of which used only the three reference sequences. First, we used parsimony, implemented in custom scripts in R. Second, we used the non-homogeneous general time-reversible (GTR-NH_b) substitution model, implemented in the *baseml* package of PAML v4.8 (Yang 2007), after checking that GTR-NH_b fitted the data better than the stationary GTR model using chi-squared tests (see Results). The use of this method to reconstruct ancestral sites when nucleotide composition is non-stationary is described in Matsumoto et al. (2015), and has been shown to produce highly accurate results in the presence of non-equilibrium base composition, whereas the parsimony method is likely to be biased. Under the GTR-NH_b method, we implemented two ways of determining the ancestral state at the *ms* node, by

either using the single best reconstruction (SBR) of the ancestral sequence at the *ms* node, or by weighting the four possible nucleotides at the *ms* node by the posterior probability of each. Instead of ignoring sub-optimal reconstructions, as the parsimony and SBR methods do, the last option weights all the possible ancestral states by their respective posterior probabilities. Following Matsumoto et al. (2015), we refer to these two GTR-NH_b-based methods as 'SBR' and 'AWP' respectively. The AWP method should be more reliable than either parsimony or SBR when base composition is not at equilibrium (Matsumoto et al. 2015).

Since some of the models we used are very parameter-rich (e.g., the GTR-NH_b model has 39 parameters for three species, and the M1* model described more fully below has 25 parameters for *D. simulans* and 21 parameters for *D. melanogaster*, given the sample sizes), we had to group genes into bins to avoid overfitting. To investigate the relationship between selection and GC content at 4-fold sites (a proxy for the extent of CUB), we binned 4-fold sites by the GC content in the *D. melanogaster* reference sequence, which we used as a proxy for the historic strength of selection favouring GC alleles. GC content evolves very slowly over time (Marais et al. 2004), and is highly correlated between *D. simulans* and *D. melanogaster* CDS (Pearson's correlation coefficient $r = 0.97$, $p < 2.2 \times 10^{-16}$), so this strategy should accurately represent GC content at the *ms* node. We binned 4-fold degenerate sites into 20 autosomal and four X-linked bins. Bins were chosen to maintain approximately the same number of genes per bin. The autosomal and X-linked SI sites were always treated as two separate bins. We also followed this binning convention for other analyses. When carrying out correlation analyses between GC content bins and other variables (e.g., substitution rate and estimates of the selection coefficient), we included only the 4-fold degenerate site GC bins, but not the SI bin. We also restricted the correlation analysis to the autosomal bins only. Given the small number of bins on the X chromosome, this type of

analysis is underpowered for the X; in fact, the smallest *p*-value that Kendall's τ can achieve with four data points is 0.08.

To determine whether or not *D. melanogaster* and *D. simulans* are in base composition equilibrium, for each bin we counted the numbers of $S \rightarrow W$ ($N_{S \rightarrow W}$), $W \rightarrow S$ ($N_{W \rightarrow S}$), and putatively neutral (N_{neu}) substitutions (i.e., $S \rightarrow S$ and $W \rightarrow W$), where S represents G or C, the strong (potentially preferred) allele, and W represents A or T, the weak (potentially unpreferred) allele. We did this along each of the *D. melanogaster* and *D. simulans* lineages by (probabilistically) comparing the reconstructed ancestral states at the *ms* node with the reference genomes. This is reasonable because the branch length is much higher than the level of within-species polymorphism (see Results). For the AWP method, we rounded our results to the nearest integer. Where possible, we compared our results to those published in the literature, and to equivalent results kindly provided by Juraj Bergman and Claus Vogl (pers. comm.; Table S2). To obtain the $W \rightarrow S$ substitution rate ($r_{W \rightarrow S}$) per bin, we divided $N_{W \rightarrow S}$ by the total number of AT sites (L_W) at the *ms* node in that bin. Similarly, $r_{S \rightarrow W} = N_{S \rightarrow W}/L_S$.

Polymorphism-based analyses

For each bin, we estimated the derived allele frequency at segregating sites, using the three methods described above to infer ancestral states at the *ms* node, which should be a reasonable approximation given the rarity of shared polymorphism for the two species (Clemente & Vogl 2012b). We classified these sites into segregating sites at which the ancestral allele was AT and the derived allele was GC ($DAF_{W \rightarrow S}$), and segregating sites at which the ancestral allele was GC and the derived allele was AT ($DAF_{S \rightarrow W}$), as well as segregating sites which had mutated from A to T, or *vice versa*, and from G to C or *vice versa*

(DAF_{neu}). We also calculated Δ_π (Langley et al. 2014) for each bin. We mostly display results obtained from the AWP method in the Results section, because it is probably the most reliable of the three. Qualitatively, the results are generally insensitive to the choice of method for reconstructing ancestral sites. Thus, we present a set of figures in the supplement (Supplementary figures S6 – S11) that are parallel to those shown in the main text, but were obtained using either parsimony or SBR, respectively.

We used two polymorphism-based methods for estimating the population-scaled strength of the force favouring GC alleles, $\gamma = 4N_e s$, where N_e is the effective population size and *s* is the selection coefficient against heterozygous carriers of the AT allele. The first is the method of Glémin et al. (2015), which uses three different classes of polarised unfolded site frequency spectra (SFS) for sites which are segregating in the present day: $S \rightarrow W$, $W \rightarrow S$, and putatively neutral (see above). This method is capable of taking into account polarization errors, which, if untreated, may lead to upwardly biases 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 (after Eyre-Walker et al. 2006). Because it only considers the SFS of derived alleles, we expect this method to recover signatures of selection on a relatively recent time scale ($\sim 4N_e$ generations, if we conservatively assume neutrality). We generated unfolded SFSs for this model using the AWP method to infer the ancestral state at the *ms* node, and estimated the strength of γ using R code provided in the supplementary material of Glémin et al. (2015). We refer to the models using this method with the same notation as Glémin et al. (2015). These are: model M0, where $\gamma = 0$ and polarisation errors are not taken into account; M1, where $\gamma \neq 0$ and polarisation errors are not taken into account; and M0* and M1*, which are the equivalent

models after correcting for polarisation errors. Note that the method for controlling for demography drastically increases the number model parameters. For instance, for M1, in addition to γ and the three mutational parameters for each of the three SFSes ($\theta = 4N_e\mu$), it requires an additional $n - 2$ nuisance parameters, where n is the number of frequency classes (in our case, this is the same as the sample size). Given the dearth of SNPs relative to substitutions, and in particular the lower diversity level in *D. melanogaster*, we repeated some of these analyses by pooling SNP data across several nearby GC content bins (see Results).

Second, we used the method of Zeng and Charlesworth (2009), modified as described by Evans et al. (2014), which uses the unpolarised SFS (including fixed sites) to infer parameters of a two-allele model with reversible mutation between W and S alleles, selection and/or gBGC, and changes in population size (see Zeng (2012) for a discussion of the differences between the reversible mutation model and the infinite-sites model on which the method of Glémin et al. (2015) is based). Because this method uses the unpolarised SFS, no outgroup is required. This method can recover signals of selection (and other population genetic parameters) over a longer time scale than the methods of Glémin et al. (2015), because it uses information on the base composition of the species to estimate the parameters (see Zeng & Charlesworth 2009; supplementary Figures S8-S11). As above, we defined W (AT) and S (GC) as our two alleles. We define u as the rate at which S alleles mutate to W alleles, and v as the mutation rate in the opposite direction, and $\kappa = u/v$ as the mutation bias parameter. To incorporate a change in population size, we assume that the population in the past is at equilibrium with population size N_1, which then changes instantaneously to N_0 (this can be either an increase or a reduction in size) and remains in this state for t generations until a sample is taken from the population in the present day (Zeng & Charlesworth 2009;

Haddrill et al. 2011; Evans et al. 2014). As with M1* and M1, we also tested the equivalent models where $\gamma = 0$. For each model, in order to ensure that the true MLE was found, we ran the search algorithm multiple times (typically 500), each initialised from a random starting point. All the results reported above were found by multiple searches with different starting conditions. Chi-squared tests were used to evaluate statistical support for different models. We refer to these models as ZC0 ($\gamma = 0$) and ZC1 ($\gamma \neq 0$) below. A software package implementing this approach is available at http://zeng-lab.group.shef.ac.uk. For all methods (Zeng & Charlesworth 2009; Glémin et al. 2015), we fitted independent models for each SI and 4-fold bin (Zeng & Charlesworth 2010b; Messer & Petrov 2013).

Results

Patterns of polymorphism and divergence in the *D. simulans* and *D. melanogaster* datasets

For *D. simulans*, after extracting 4-fold degenerate sites and short introns (SI; positions 8-30bp of introns <66bp long), we retained 7551 autosomal coding sequences (CDS) alignments and 1226 X-linked CDS alignments, as well as 5578 autosomal SI alignments and 516 X-linked SI alignments. The final dataset contained the reference sequences of *D. simulans, D. melanogaster* and *D. yakuba*, as well as polymorphism data from 39 *D. simulans* lines, including 21 Madagascan (MD) lines and 18 Kenyan (NS) lines, with 22 of the 39 lines being described for the first time in this paper (see Material and Methods). For *D. melanogaster*, we retained 5550 autosomal CDS alignments and 888 X-linked CDS alignments, as well as 7397 autosomal SI alignments and 738 X-linked SI alignments, containing polymorphism data from 17 Rwandan (RG) lines, as well as the three reference sequences.

Summary statistics calculated using a *D. simulans* dataset that was filtered to separate true genetic variation from variant-calling artefacts are presented in Table 1 (see Table S1 for the unfiltered data). Consider first the MD lines (n = 21) collected from the putatively ancestral range of the species in Madagascar (Dean & Ballard 2004). Autosomal π at 4-fold sites (referred to as π_4) was 0.0329 and 0.0317 for the unfiltered and filtered datasets, respectively, similar to the value of 0.035 reported by Begun et al. (2007). On the X, π_4 was 0.0191 and 0.0182 for the two datasets; the Begun et al. (2007) value was 0.02. Tajima's D and Δ_π at 4-fold sites are both negative, implying that there may have been a substantial recent population size expansion. Again, values obtained from the filtered and unfiltered data are very similar (cf. Tables 1 and S1). Overall, diversity was slightly reduced for our filtered dataset, which

may have been a result of more conservative variant filtering criteria, but the differences are minimal. In what follows, we only present results obtained from the filtered dataset. SI sites, which we only obtained from our filtered dataset, are more diverse than 0-fold and 4-fold sites in the MD population, for both the autosomes (A) (π_{SI} = 0.0321) and the X (π_{SI} = 0.0208) (Table 1).

The samples collected from Kenya (the NS lines; n = 18) have consistently lower diversity levels at 0-fold, 4-fold and SI sites, and less negative Tajima's D and Δ_π, probably caused by bottlenecks associated with the colonisation process (Dean & Ballard 2004). Nonetheless, F_{ST} between the two populations at 4-fold sites is rather low: ~2.5% between NS and MD (Table 1), suggesting that there is relatively little genetic differentiation between the ancestral and derived populations. There is also little difference in F_{ST} at 4-fold sites between the X and A. Similar to the MD population, SI sites are the most diverse class of site as measured by π (Table 1).

The patterns reported above contrast with those observed in *D. melanogaster* (see Table 1 of Jackson et al. (2015)). We focus first on samples from the putatively ancestral ranges of both species (i.e., the Rwandan (RG) lines for *D. melanogaster,* and the MD lines for *D. simulans*). Autosomal π_4 is ~2.06 times higher in *D. simulans*, suggestive of higher N_e, which may lead to more effective selection (see Discussion). Tajima's D is also less negative in *D. melanogaster*, with the differences at 4-fold sites being the most noticeable (-0.11 *vs* -1.03 for A, and -0.47 *vs* -1.31 for the X), suggesting a more stable recent population size in *D. melanogaster*, which is supported by the fits of the Zeng and Charlesworth (ZC) method to the data (see below). The X:A ratio of π_4 in *D. melanogaster* was 1.08, much higher than the expected value of 0.75 under the standard neutral model, whereas it was 0.57 in *D. simulans*.

Furthermore, F_{ST} at 4-fold sites between RG and a sample from France (Jackson et al. 2015) in *D. melanogaster* is ~10 times higher than that between the MD and NS populations in *D. simulans*. Interestingly, the difference in F_{ST} between the X and A is much more marked in *D. melanogaster* (0.29 vs. 0.17 for the X and A, respectively) than in *D. simulans* (0.025 for both X and A). Various theories have been proposed to explain differences in diversity levels between X and A, which include sex-specific variance in reproductive success (Charlesworth 2001), demographic effects (Pool & Nielsen 2007; Singh et al. 2007; Pool & Nielsen 2008; Yukilevich et al. 2010), positive and negative selection (Singh et al. 2007; Charlesworth 2012), and differences in recombination rate (Charlesworth 2012). Detailed analyses of the factors underlying X-autosomal differences are outside the scope of this study; below we present results from X and the autosomes separately.

We also assayed divergence between the reference sequences in our alignments. Between *D. melanogaster* and *D. simulans*, K_A, K_S and K_{SI} were 0.014, 0.109 and 0.130, respectively. These values are similar to those in Table 1 of Parsch et al. (2010) (K_A = 0.019, K_S = 0.106 and K_{SI} = 0.123), and in Zhang et al. (2013; Table S2) (K_A = 0.015 and K_S = 0.12). In our data K_A, K_S and K_{SI}, between *D. melanogaster* and *D. yakuba* were 0.036, 0.266 and 0.294, respectively; between *D. simulans* and *D. yakuba*, they were 0.036, 0.250 and 0.302, respectively. Note that divergence is always highest at the SI class of site, which is in agreement with these sites being relatively unconstrained (Halligan & Keightley 2006; Parsch et al. 2010; Clemente & Vogl 2012b). Overall, these patterns suggest that our alignments are of high quality.

In the following sections of this paper, we first focus on analysing the forces that act on 4-fold sites. To investigate the relationship between selection and GC content at 4-fold sites (a

proxy for the extent of CUB), we binned 4-fold sites by their GC content in the *D. melanogaster* reference sequence, which we used as a proxy for the historic strength of selection favouring GC alleles. In this part of the analysis, the putatively neutrally-evolving SI sites are analysed as a whole and presented alongside results from 4-fold sites for comparison. Later, to gain further insights into the evolution of the SI sites themselves, we binned them according their GC content, and analysed the bins in the same manner as the 4-fold sites. Only data from the putatively ancestral populations (i.e., MD in *D. simulans* and RG in *D. melanogaster*) are considered, in order to avoid complications introduced by population structure. For ease of notation, we use GC and S (the strong, potentially preferred allele) interchangeably below; the same applies to AT and W (the weak, potentially unpreferred allele).

Excess of $S \rightarrow W$ substitutions at 4-fold sites on both the *D. simulans* and the *D. melanogaster* lineages

For all the 4-fold site bins and the SI bin (on both A and X), a non-homogeneous (GTR-NH_b) substitution model implemented in PAML always fitted the data significantly better than a stationary (GTR) substitution model in both species (min $\chi^2 = 166.86$, d.f. = 28, $p = 1.05 \times 10^{-21}$), which is indicative of a non-equilibrium base composition. Considering the genome as a whole, both the *D. melanogaster* and *D. simulans* lineages showed an excess of $S \rightarrow W$ changes at autosomal and X-linked 4-fold degenerate sites, regardless of which method was employed to infer ancestral states at the *melanogaster-simulans* (*ms*) node (Tables 2 and S2; see Material and Methods). It is evident that the excess is greater in *D. melanogaster* than *D. simulans*. For instance, based on autosomal data obtained by the AWP method, which we expect to be the most accurate method of the three (Matsumoto et al. 2015), the ratio $N_{W\rightarrow S}/N_{S\rightarrow W}$, where $N_{W\rightarrow S}$ and $N_{S\rightarrow W}$ are the numbers of substitutions between the S and W

alleles along the lineage of interest, is 0.49 in *D. simulans*, but is only 0.26 in *D.* melanogaster (χ^2 = 2145.8, d.f. = 1, $p < 0.001$). Interestingly, the $S \rightarrow W$ bias is much more pronounced on the X of *D. melanogaster* with an $N_{W\rightarrow S}/N_{S\rightarrow W}$ ratio of 0.17, significantly different from the A value of 0.26 (χ^2 = 212.8, d.f. = 1, $p < 0.001$), whereas in *D. simulans* the ratios are much closer to one another, 0.53 and 0.49, respectively, although this difference is still significant (χ^2 = 6.97, d.f. = 1, $p = 0.008$). These results are in line with previous findings of an excess of AT (or unpreferred codon) substitutions at silent sites in *D. melanogaster* (Akashi 1995, 1996; Takano-Shimizu 2001; Akashi et al. 2006). For *D. simulans*, our data are in agreement with a dataset curated entirely independently by Juraj Bergman and Claus Vogl (pers. comm.; Table S2), and suggest that there is a much more pronounced $S \rightarrow W$ bias than was found in some previous studies (Akashi et al. 2006; Begun et al. 2007; Poh et al. 2012).

The ratio $N_{W\rightarrow S}/N_{S\rightarrow W}$ is much closer to unity for SI sites than for 4-fold sites (Table 2), which is also in agreement with the previous finding that short introns are generally closer to equilibrium than 4-fold sites in both species (Kern & Begun 2005; Singh et al. 2009; Haddrill & Charlesworth 2008; Robinson et al. 2014). The three methods for inferring ancestral states in the *ms* ancestor consistently suggest an AT substitution bias at SI sites in the *D. melanogaster* lineage (Table 2). The situation is somewhat more complex in *D. simulans.* For the X, all three methods suggest a mild GC bias, but the ratio based on AWP, which should be the most reliable method of the three (Matsumoto et al. 2015), is not significantly different from 1 (χ^2 = 0.286, d.f. = 1, $p = 0.59$). For the autosomes, parsimony suggests a GC bias (χ^2 = 19.7, d.f. = 1, $p = 0.01$), but both SBR and AWP provide some support for a slight AT bias (SBR: χ^2 = 3.73, d.f. = 1, $p = 0.05$; AWP: χ^2 = 5.55, d.f. = 1, $p = 0.019$) (Table 2). This may

reflect the tendency for parsimony to overestimate changes from common to rare basepairs (Collins et al. 1994; Eyre-Walker 1998; Akashi et al. 2007; Matsumoto et al. 2015).

Variation in 4-fold site substitution patterns across regions with different GC content

Under strict neutrality, the substitution rate per site is equal to the mutation rate per site (Kimura 1983). Thus, if 4-fold degenerate sites have never been affected by selection on CUB and/or gBGC, the two substitution rates per site, $r_{W\to S}$ and $r_{S\to W}$, should be uniform across the GC bins, unless there are systematic differences in mutation rates across bins. However, as can be seen from Figure 1, in both species, on both the autosomes and the X chromosome, $r_{W\to S}$ is positively correlated with GC content (*D. simulans*, autosomes: Kendall's $\tau = 0.45$, $p = 0.006$; *D. melanogaster*, autosomes: $\tau = 0.53$, $p = 0.001$). Here and in what follows, we refrain from conducting formal correlation tests of the X-linked data due to the dearth of data points; in addition, data from the SI bins is not included in correlations. In contrast, $r_{S\to W}$ shows a clearly negative relationship with GC content (Kendall's $\tau = -0.95$, $p < 0.001$ and $\tau = -0.96$, $p < 0.001$ for *D. simulans* and *D. melanogaster* autosomes, respectively). These patterns are expected if GC alleles (i.e., preferred codons) were favoured over AT alleles (i.e., unpreferred codons) for a substantial amount of time along these two lineages, and the intensity of the GC-favouring force increases with GC content (see the Discussion for an explicit model). Also of note is the marked increase in $r_{S\to W}$ relative to $r_{W\to S}$ with GC content in the *D. melanogaster* lineage, which is suggestive of mutations becoming more AT-biased. However, the arguments set out in the Discussion suggest that a change in mutational bias alone is unlikely to explain the data reported here.

As stated before, the $N_{W\to S}/N_{S\to W}$ ratio at SI sites, particularly in *D. simulans*, is close to unity, the value expected under equilibrium base composition. An investigation across the 4-

fold site GC content bins suggests that all of the bins considered here are experiencing some level of AT fixation bias ($N_{W\to S}/N_{S\to W} < 1$), and that genomic regions with higher GC contents are evolving towards AT faster than regions with lower GC contents. This is clear from the negative correlations between GC content and the level of substitution bias ($N_{W\to S}/N_{S\to W}$) calculated per 4-fold site bin in both species (Kendall's τ = -0.96, $p < 0.001$ and τ = -0.91, $p < 0.001$ for *D. simulans* and *D. melanogaster* autosomes, respectively) (Figure 2). As explained in the Discussion, this negative correlation can readily be explained by a genome-wide reduction in the intensity of the GC-favouring force.

Derived allele frequencies (DAF) at 4-fold sites provide clear evidence of ongoing selection for preferred codons

If selection/gBGC favours GC alleles over AT alleles, then the frequencies of derived GC alleles at AT/GC polymorphic sites ($DAF_{W\to S}$) should on average be higher than the frequencies of derived AT alleles at AT/GC polymorphic sites ($DAF_{S\to W}$). Furthermore, $DAF_{W\to S}$ should increase as the GC-favouring force becomes stronger (i.e., as 4-fold site GC content increases), whereas $DAF_{S\to W}$ should decrease with increasing GC content. In addition, we expect DAF_{neu}, the DAF for putatively neutral changes (i.e., segregating sites which had mutated from A to T, or vice versa, and from G to C or vice versa), to lie in a position intermediate between $DAF_{S\to W}$ and $DAF_{W\to S}$ (i.e., $DAF_{W\to S} > DAF_{neu} > DAF_{S\to W}$). In contrast, in a neutral model with a recent increase in mutational bias towards AT, the higher number of derived AT mutations entering the population, which tend to be young and segregate at low frequencies, will depress $DAF_{S\to W}$, leading to $DAF_{W\to S} > DAF_{S\to W}$, but DAF_{neu} should be comparable to $DAF_{W\to S}$. Moreover, GC content and $DAF_{W\to S}$ should be unrelated under this model.

D. simulans fits the expectations of the first model: $DAF_{W\to S}$ is greater than $DAF_{S\to W}$ in all autosomal and X-linked 4-fold bins, and DAF_{neu} is always intermediate between $DAF_{W\to S}$ and $DAF_{S\to W}$ (Figure 3). Autosomal 4-fold site $DAF_{W\to S}$ correlates positively with GC content (Kendall's τ = 0.6, $p < 0.001$; Figure 3), and autosomal 4-fold site $DAF_{S\to W}$ correlates negatively with GC content (Kendall's τ = -0.85, $p < 0.001$; Figure 3); data from the X display similar trends. These patterns suggest the action of forces favouring GC over AT alleles in the recent past in this species (a time period of the order of $4N_e$ generations), with higher GC content bins experiencing a higher strength of recent selection favouring GC.

In *D. melanogaster*, the equivalent results are less clear. Autosomal $DAF_{W\to S}$ is higher than autosomal $DAF_{S\to W}$ for 19/20 4-fold bins (Figure 3). As in *D. simulans*, autosomal 4-fold $DAF_{W\to S}$ correlates positively with GC content (Kendall's τ = 0.41, p = 0.01; Figure 3), and autosomal 4-fold $DAF_{S\to W}$ correlates negatively with GC content (Kendall's τ = -0.47, p = 0.004; Figure 3). DAF_{neu} falls between $DAF_{W\to S}$ and $DAF_{S\to W}$ in 14/20 autosomal 4-fold site bins, but only 1/4 X-linked 4-fold bins (Figure 3). Additionally, the difference between $DAF_{W\to S}$ and $DAF_{S\to W}$ seems less pronounced than in *D. simulans*, especially on the X chromosome, although on the autosomes the gap between $DAF_{W\to S}$ and $DAF_{S\to W}$ does tend to increase with GC content and is the largest and most comparable in magnitude to those seen in *D. simulans* in the bins with the highest GC content. Overall, these data provide some evidence of recent selection for GC at 4-fold sites in *D. melanogaster*, but its extent seems to be smaller than in *D. simulans*, and may be restricted to autosomal regions with high GC contents.

Estimating γ and other parameters using 4-fold site polymorphism data

To shed further light on the evolutionary dynamics of selection on CUB, we used two different methods for inferring the scaled strength of selection for GC alleles ($\gamma = 4N_e s$) from polymorphism data. First, we applied the method of Glémin et al. (2015), which detects recent selection (timescale $\sim 4N_e$ generations). We refer to the different variants of this method using the same notation as Glémin et al. (2015). These are: model M0, where $\gamma = 0$ and polarisation errors (with respect to inferring ancestral vs. derived alleles) are not taken into account; M1, where $\gamma \neq 0$ and polarisation errors are not taken into account; and M0* and M1*, which are the equivalent models after correcting for polarisation errors. Second, we used the method of Zeng and Charlesworth (2009), modified as described by Evans et al. (2014), which provides estimates over a longer period. We used two variants of this method, which are referred to as ZC0 ($\gamma = 0$) and ZC1 ($\gamma \neq 0$).

For every *D. simulans* bin on both the A and X, both ZC1 and M1 fit the data significantly better than the corresponding models with $\gamma = 0$ (i.e., ZC0 and M0; min χ^2 = 17.84, d.f. = 1, p < 0.001); the only exception is the X-linked SI bin where M1 does not fit the data better than M0 (χ^2 = 0.071, d.f. = 1, p = 0.79) (Figure 4). Estimates obtained by ZC1 and M1 agree closely for the *D. simulans* data (Figure 4; Wilcoxon paired signed rank test, p = 0.25). The agreement between the results from the two methods, which are expected to be sensitive to forces favouring GC on different timescales (see Material and Methods), suggests consistent selection over time favouring GC alleles at 4-fold degenerate sites in *D. simulans*. In addition, GC content correlates positively with γ on both the autosomes (Kendall's τ = 0.98, p < 0.001; τ = 0.88, p < 0.001 for ZC1 and M1, respectively) and the X chromosome. Thus, in agreement with the results obtained from the divergence- and DAF-based analyses, selection for GC is indeed stronger in regions with higher GC content. The patterns obtained

from comparing M0* and M1* are qualitatively identical (Supplementary figure S2). In addition, when using the Akaike Information Criterion (AIC) to rank the four Glémin models (this is necessary because, e.g., M0* and M1 are not nested and cannot be compared using the likelihood ratio test), M1 and M1* are always the two best fitting models for all bins across both chromosome sets, except for the SI bin on the X (Table S3).

Similarly to the analysis based on DAFs, the patterns are less clear-cut in *D. melanogaster*. When M1 and M0 are compared, 13/20 autosomal 4-fold site bins are found to be non-neutrally evolving, including the four highest autosomal GC bins, and none on the X (Figure 4). In contrast, according to the comparison between M1* and M0*, only 3 autosomal bins show evidence of non-zero γ in *D. melanogaster* (2/20 autosomal 4-fold site bins and the autosomal SI bin), and none of the X-linked bins do so (Supplementary figure S2). In particular, the fact that none of the high GC bins have a significant test is out of keeping with the observation that these bins have large differences between $DAF_{W\to S}$ and $DAF_{S\to W}$. A close inspection suggests that statistical power may be an issue: there are on average four times fewer SNPs in the 4-fold site bins in *D. melanogaster*, and in the highest 4-fold site bin, there were only 69 $W \to S$ SNPs. As described in the Material and Methods, the Glémin models are parameter-rich, especially M0* and M1*. In fact, M1* often came out (e.g., in 10/20 autosomal 4-fold site bins) as the worse fitting one among the four models according to the AIC.

To deal with this issue, we redid the comparison by reducing the number of autosomal 4-fold bins to 10. M1 fits better than M0 in 9/10 bins, while M1* fits better than M0* in 4/10 bins, including two out of the top four GC bins (Supplementary figure S3). According to the AIC, the frequency of M1 being the best fitting model increases to 9/10 bins, whereas the

frequency of M1* being the worse fitting model decreases to 2/10 bins (Table S3). The observation that M1* sometimes ranked lower than M1 according to the AIC in both species may also be due to the fact that our method for correcting for non-equilibrium when reconstructing ancestral states has reduced the need to correct for polarisation errors.

As is apparent from Figure 4, M1 also estimates consistently lower absolute values of γ than ZC1 in *D. melanogaster* (Wilcoxon paired signed rank test, $p = 1.9 \times 10^{-6}$). Given that the ZC method returns long-term average estimates of γ, these differences clearly indicate a recent decline in the strength of selection on CUB in this species. As with *D. simulans*, however, autosomal GC content correlates positively with γ under both models (Kendall's $\tau = 0.87$, $p < 0.001$; $\tau = 0.48$, $p = 0.003$ for ZC and M1, respectively; Figure 4), which is suggestive of some, if weak, ongoing selection for GC at autosomal 4-fold sites, particularly in GC-rich regions of the genome. The fact that the SFS is more negatively skewed at 4-fold sites in regions of higher GC content in both species, as measured by Δ_π (Supplementary figure S4), is also consistent with selection on these sites.

In addition to γ, the two methods also produced estimates of other parameters of interest. For instance, both methods can estimate κ, the mutational bias parameter, defined as u/v where u is the mutation rate from *S* to *W* per site per generation, and v is that in the opposite direction. As shown in Supplementary Figure S12, in *D. simulans*, κ is close to 2 across the 4-fold site bins, similar to previous estimates obtained by different methods (Singh et al. 2005; Keightley et al. 2009; Zeng 2010; Schrider et al. 2013). The fact that κ is estimated to be similar across the bins suggests that the difference in 4-fold sites' GC content can be attributed to stronger selection, not to differences in mutational bias. In *D. melanogaster*, the difference in the estimates between the two methods is much more pronounced, with κ from

the Glémin method (short timescale) being consistently higher than those estimated by the ZC method (long timescale), probably reflecting a recent increase in the mutation rate towards A/T nucleotides (see Discussion).

Consistent with the apparently negative Tajima's *D* values calculated using 4-fold sites in *D. simulans* (Table 1), the ZC method detected clear evidence for recent population expansion in all bins ($p < 10^{-16}$ for all bins; Supplementary Table S4), whereas for *D. melanogaster*, no clear evidence for recent population expansion was found, which is consistent with the observed data (e.g., Tajima's *D* is only -0.11 for A in *D. melanogaster*, but is -1.03 in *D. simulans*) and our previous analysis based on a different dataset (Zeng & Charlesworth 2009). In Supplementary Text S2 (see also Supplementary Tables S5 and S6), we present a more detailed description of estimation of the demographic parameters in *D. melanogaster*, and the statistical and computational issues we encountered. We also provide evidence that our conclusion of a continuing decline in selection intensity in *D. melanogaster* is robust to these potential issues (Supplementary Figure S13).

A more detailed analysis of the short introns

The SI data shown in Figures 3 – 4 suggest that GC may be favoured over AT in short introns. Given the apparent lack of selective constraints on SI sites (Halligan & Keightley 2006; Parsch et al. 2010), this is suggestive of the action of gBGC. In contrast to selection on CUB at 4-fold sites, all alleles have equal fitness under the gBGC model, and the selection-like pattern is created by the preferential transmission of the *S* allele in *SW* heterozygotes to the next generation (Duret and Galtier 2009). The $S \rightarrow S$ and $W \rightarrow W$ mutations are "neutral" in the sense that they should be unaffected by gBGC. To gain further insights, we carried out additional analyses by binning the SI data according to their GC content, and asked whether

gBGC could be responsible for the observed patterns. Constrained by the limited amount of data and the parameter-richness of some of the models, we only carried out these analyses using the autosomal SI data, divided into 5 bins. These data were then examined in the same way as the 4-fold sites. However, with such a small number of bins, the correlation-based analysis is likely to be prone to statistical noise; the results should thus be treated with caution.

As shown in Figures 5A and 5E, $r_{S\to W}$ decreases as GC content increases in both species (Kendall's τ = -1, p = 0.03), which may reflect an ancestral reduction in the strength of the force favouring G/C nucleotides (see Discussion). However, $r_{W\to S}$ is not significantly correlated with GC content in either species (Kendall's τ = -0.8, p = 0.09, in *D. simulans*; Kendall's τ = 0.8, p = 0.09, in *D. melanogaster*). Comparing $N_{W\to S}$ and $N_{S\to W}$ across bins using a 2 x 5 contingency table test suggests that the substitution pattern is heterogeneous across the bins in both species ($p < 2.2 \times 10^{-16}$ in *D. simulans* and $p = 2.04 \times 10^{-8}$ in *D. melanogaster*). The $N_{W\to S}/N_{S\to W}$ ratio decreases with increasing GC content in *D. simulans* (Kendall's τ = -1, p = 0.03; Figure 5B), qualitatively similar to what we reported above for the 4-fold sites in this species (Figure 2). However, this ratio shows no significant correlation with GC content in *D. melanogaster* (Kendall's τ = 0.8, p = 0.09; Figure 5F). These results highlight the difficulty in conducting detailed analyses in the SI regions, due to insufficient data. Nevertheless, they provide evidence for variation between different SI regions.

We did not detect any statistically significant correlation between the three types of DAFs and GC content in *D. simulans* (Figure 5C, minimum p = 0.22 for the three tests), although the relationship $DAF_{S\to W} < DAF_{neu} < DAF_{W\to S}$ holds in all bins. The lack of strong support for a relationship with GC content was also reflected when the Kruskal-Wallis test was used

to test for heterogeneity in median DAFs across bins; the p-values for $S \rightarrow W$, neutral, and $W \rightarrow S$ are 0.38, 0.20 and 0.04, respectively. In *D. melanogaster* (Figure 5G), $DAF_{S \rightarrow W}$ is significantly negatively correlated with GC content (Kendall's $\tau = 1$, $p = 0.03$), but no relationship was found for the other two DAFs (minimum $p = 0.22$). In the three bins with higher GC content, we have $DAF_{S \rightarrow W} < DAF_{neu} < DAF_{W \rightarrow S}$. But the order is completely reversed in the lowest GC content bin, although the differences between the DAFs are non-significant based on the Glémin model (see below). Consistent with this, the Kruskal-Wallis test detected significant heterogeneity in median DAF across bins in the $DAF_{S \rightarrow W}$ case ($p = 1.40 \times 10^{-8}$), but not in the other two cases ($p > 0.08$).

Finally, we used polymorphism data to estimate the strength of the force favouring GC, as measured by γ. In line with the DAF-based analysis, in neither *D. simulans* (Kendall's $\tau = 0$, $p = 1$; Figure 5D) nor *D. melanogaster* (Kendall's $\tau = 0.8$, $p = 0.09$; Figure 5H) did we find a significant relationship between GC content and γ as estimated by the M1 model of Glémin et al. (2015). In *D. simulans*, M1 fits the data significantly better than M0 in all five bins, whereas in *D. melanogaster*, the neutral model M0 is sufficient to explain the data for the first two bins, with the M1 model being more adequate for data collected from the more GC-rich bins. Estimates of γ produced by the ZC1 method are positively correlated with GC content in both species (Kendall's $\tau = 1$, $p = 0.03$; Figures 5D and 5H). Interestingly, ZC1 fits the data significantly better than ZC0 in all cases, even in bins where γ is fairly close to zero. A close inspection suggests that this is not due to poor convergence in the search algorithm. Furthermore, simulations have shown that the ZC model is very robust to linkage between sites and demographic changes (Zeng & Charlesworth 2010b), suggesting that these results are unlikely to be methodological artefacts, and may reflect long-term dynamics in these regions. Finally, in *D. melanogaster*, there is no clear evidence that the estimates of long-

term γ derived from ZC1 are higher than estimates of short-term γ derived from M1 (Figure 5H).

Discussion

Evidence for past selection on CUB in both *Drosophila* species

The correlations between the substitution rates and GC content at 4-fold sites presented in Figures 1 and S5 can be explored using the following modelling framework (Li 1987; Bulmer 1991; McVean & Charlesworth 1999), which assumes a fixed N_e and thus a fixed value of γ for each GC bin. If there are temporal changes along a lineage, we can regard these parameters as long-term averages. Let u be the mutation rate from $S \rightarrow W$ per site per generation; and v be that in the opposite direction. Define κ as u/v. The two substitution rates, $r_{S\rightarrow W}$ and $r_{W\rightarrow S}$, are proportional to $u\gamma/[exp(\gamma)-1]$ and $v\gamma/[1-exp(-\gamma)]$, respectively (e.g., Eq. B6.4.2b of Charlesworth and Charlesworth (2010); Eq. 11 of Sawyer and Hartl (1992); Akashi et al. 2007). We can then define

$$R = \frac{r_{S\rightarrow W}}{r_{W\rightarrow S}} = \kappa \frac{1-e^{-\gamma}}{e^{\gamma}-1} = \kappa e^{-\gamma} \quad (2)$$

Assuming that u and v are constant across the GC bins and over time (κ is thus also constant), R is a function of γ. Taking the derivative with respect to γ, we have

$$\frac{dR}{d\gamma} = -\kappa e^{-\gamma} \quad (3)$$

In other words, $R = \kappa$ when $\gamma = 0$ (neutrality), and decreases as γ becomes positive (i.e., when W is selected against). Thus, the decreasing values of R shown in Figures 1 and S5 suggest that S is more strongly favoured in high GC bins. For instance, the R values for the

lowest and highest autosomal 4-fold site bins in *D. simulans* are 1.51 and 0.56, respectively. If the SI sites are neutral (see below), κ can be estimated by the R value from the SI bin, which is 1.93, very close to the value of 2 reported previously (Singh et al. 2005; Keightley et al. 2009; Zeng 2010; Schrider et al. 2013), solving Eq. (2) for γ gives values of 0.25 and 1.24 for the lowest and highest bins, respectively. These rough, long-term estimates are about 2-fold lower than those obtained from the polymorphism data (Figure 4). It is possible that *D. simulans* has a larger recent N_e (reflected in the polymorphism-based analysis) than the average N_e along the entire lineage, which is consistent with the evidence for population expansion from the negative Tajima's *D* values (Table 1). Finally, as detailed in the Supplementary text S1, this model can also explain why the slope for $r_{S\to W}$ is apparently steeper than that for $r_{W\to S}$ (Figure 1).

The above model can also explain why, at 4-fold sites, $R_N = N_{W\to S}/N_{S\to W} < 1$ and there is a negative relationship between R_N and GC content (Figure 2), where $N_{W\to S}$ and $N_{S\to W}$ are the numbers of substitutions between the S and W alleles along the lineage of interest. Note first that $N_{S\to W}$ and $N_{W\to S}$ are, respectively, proportional to $Qu\gamma/[exp(\gamma)-1]$ and $(1-Q)v\gamma/[1-exp(-\gamma)]$, where Q is the GC content at the *ms* node (since Q changes very slowly, this should be a reasonable first approximation). At equilibrium, $Q = 1/[1+\kappa exp(-\gamma)]$ (Li 1987; Bulmer 1991) and hence $N_{W\to S}/N_{S\to W} = 1$. Consider a model where the ancestral species was at equilibrium, but γ is reduced to $p\gamma(0 \le p < 1)$ along a lineage that leads to an extant species, so that $N_{S\to W}$ and $N_{W\to S}$ become proportional to $Qup\gamma/[exp(p\gamma)-1]$ and $(1-Q)vp\gamma/[1-exp(-p\gamma)]$, respectively. Then, R_N for the GC content bin in question can be written as

$$R_N = \frac{N_{W\to S}}{N_{S\to W}} = \frac{(1-Q)(e^{p\gamma}-1)}{\kappa Q(1-e^{-p\gamma})} = e^{-(1-p)\gamma} \quad (4)$$

Assuming that p is constant across bins (i.e., there has been a genome-wide proportional reduction in γ), then R_N decreases as γ increases. This, together with the arguments presented above that the long-term average γ is higher in high GC bins, Eq. (4) implies that the negative relationship between R_N and GC content is consistent with a genome-wide reduction in the intensity of selection in both species (see also Akashi et al. 2007).

In contrast, if we assume that $\gamma = 0$ and κ is constant across the bins (i.e., there has been no selection along both the *D. melanogaster* and *D. simulans* lineages), the fact that $R = \kappa$ means that a genome-wide increase in κ (i.e., a more AT-biased mutation pattern) would not cause a negative relationship between R and GC content. If the relationship between R and GC content were entirely mutational in origin, then u must decrease as GC content increases, whereas v changes in the opposite direction (Figure 1). Such a model is incompatible with the evidence for selection from the two polymorphism-based methods (Figure 4), and cannot easily explain the well-known positive correlation between GC content of coding sequences (or the extent of CUB) and gene expression levels (e.g., Campos et al. 2013), especially when considering the lack of support for transcription-coupled mutational repair in *Drosophila* (Singh et al. 2005; Keightley et al. 2009).

As shown in Supplementary Figure S12, the Glémin method (short timescale) and the ZC method (long timescale) returned κ estimates that are more comparable in *D. simulans* than in *D. melanogaster*; the ZC method produced consistently lower estimates in *D. simulans* and consistently higher estimates in *D. melanogaster* (two-sided binomial test, $p = 1.91 \times 10^{-6}$ in both cases). Taken at face value, these results suggest that there probably has been relatively

little change in the extent of mutational bias in the *D. simulans* lineage, whereas mutation may have become more AT-biased in *D. melanogaster*. These results suggest that the patterns shown in Figure 2 are probably a result of an ancestral reduction in the efficiency of selection in *D. simulans*. For *D. melanogaster*, it is possible that a more AT-biased mutational pattern has also contributed to the evolution of base composition in its genome, as suggested by previous studies (Takano-Shimizu 2001; Kern & Begun 2005; Nielsen et al. 2007; Zeng & Charlesworth 2010a; Clemente & Vogl 2012b).

Overall, the above considerations suggest that the data presented in Figures 1, 2 and S5 cannot be explained by a shift towards a more AT-biased mutational pattern alone. Instead, selection favouring GC over AT basepairs must have acted on both species for a significant amount of time since they last shared a common ancestor, although both lineages are likely to have experienced an ancestral reduction in the efficacy of selection that led to the AT-biased substitution patterns.

Estimating the intensity of selection on preferred codons over different timescales

A novelty of this study is that, by applying two different methods to the same polymorphism dataset, we have attempted to understand how the selective pressure on CUB has changed over time by comparing γ estimates reflective of either a short timescale (for roughly the last $4N_e$ generations; i.e., the Glémin method (Glémin et al. 2015)), or a long timescale (for > $4N_e$ generations; i.e., the ZC method (Zeng & Charlesworth 2009)). However, pinpointing the exact timescale for the ZC method is difficult, because it depends on details of past evolutionary dynamics that we know little about (e.g., the timescale can be affected by both the time when the ancestral population size reduction took place and the severity of the reduction; see Supplementary Figures S8 – S11 in Zeng & Charlesworth (2009)). This

difference in timescale between the methods is due to the use of the derived SFS under the infinite-sites model (Kimura 1983) in the Glémin method and the use of a reversible mutation model in the ZC method (see Zeng (2012) for a more thorough discussion of the differences between these two models). By the same token, we can classify other polymorphism-based methods into short timescale (Akashi & Schaeffer 1997; Bustamante et al. 2001) and long timescale (Maside et al. 2004; Cutter & Charlesworth 2006; Galtier et al. 2006; Zeng 2010; Clemente & Vogl 2012a; Vogl & Bergman 2015).

Contrasting the results obtained from the ZC method with those from the divergence-based analysis (Figures 1 and 2) and the Glémin method (Figure 4) is informative. First, consider *D. simulans*. The fact that values of γ estimated by both the ZC method and the Glémin method are virtually identical suggests that there have not been significant changes in the intensity of selection over the time period that the ZC method considers. Hence, the reduction in γ suggested in the previous section, which may have caused $N_{W\to S}/N_{S\to W} < 1$ and the negative correlation between $N_{W\to S}/N_{S\to W}$ and GC content, probably happened so early during the evolution of *D. simulans* that it did not leave detectable traces in the polymorphism data.

In contrast, in *D. melanogaster*, both the divergence-based analysis and the comparison between the ZC method and the Glémin method provide evidence for a reduction in γ, indicating a recent decline in this species. Assuming that short introns are neutral, and using autosomal data from the putatively ancestral populations (i.e., MD and RG), Table 1 in this study and Table 1 in Jackson et al. (2015) suggest that N_e is 2.21-fold higher in *D. simulans* compared with *D. melanogaster*, implying more efficient selection in the recent past. In fact, focusing on the 13 autosomal 4-fold site bins in *D. melanogaster* where M1 fits the data better than M0 (filled squares in Figure 4), the γ estimates in the corresponding bins in *D.*

simulans are on average 2.93 times higher, comparable to the difference in N_e suggested by the short intron data. This difference in N_e may be due to differences in the two species' demographic history. Previous studies have also suggested that the lower recombination rate in *D. melanogaster* compared to *D. simulans* (Comeron et al. 2012; True et al. 1996) may have played a role through stronger Hill-Robertson interference between selected sties (Takano-Shimizu 1999; McVean & Charlesworth 2000; Comeron et al. 2008, 2012; Cutter & Payseur 2013). However, without detailed genetic maps from closely-related outgroup species, it is impossible to ascertain whether the reduced map length in *D. melanogaster* represents the ancestral or derived state; this is an important area for further research.

Comparison with previous studies

Poh et al. (2012) suggested that AT-ending codons might be favoured in *D. melanogaster*, based on the observation that, along the *D. melanogaster* lineage, $S \rightarrow W$ mutations fixed at a higher rate than $W \rightarrow S$ changes; also, in their polymorphism dataset, the proportion of singletons in the SFS for $S \rightarrow W$ changes was smaller than in the SFS for $W \rightarrow S$ changes (23.2% vs 24.3%). The latter difference is significant under a Mann-Whitney U test, although neither Tajima's *D* nor Fu and Li's *D** were significantly different from zero. Here we have provided evidence that the pattern of $r_{S\rightarrow W} > r_{W\rightarrow S}$ can be readily explained by a reduction in selection intensity favouring *S* basepairs along the *D. melanogaster* lineage. As for their polymorphism data, Poh et al. (2012) used lines collected from Raleigh, North America. There is clear evidence that this population has experienced bottlenecks in the recent past, as can be seen from the lower level of diversity in this population compared to populations from Africa (genome-wide π_S = 0.013 *vs* 0.019 for the Raleigh and Malawi populations; Langley et al. (2012)). Without using model-based methods to correct for the effects of demographic changes, the results of Poh et al. (2012) may be susceptible to complications caused by such

complex demography. In addition, their ancestral states were inferred using maximum parsimony, which is prone to error. In Supplementary Figure S14, we used parameter values realistic for *D. melanogaster* to show that, with demography and polarisation error, it is possible for the proportion of singletons in the SFS for $S \rightarrow W$ changes to be lower than that for $W \rightarrow S$ changes in the presence of weak selection favouring *S* (see the legend to Figure S14 for further discussion of this issue).

Another possible cause of the Poh et al. (2012) results is admixture with African *D. melanogaster* during the recovery from the bottleneck (Caracristi & Schlötterer 2003; Duchen et al. 2013; Bergland et al. 2016). Because the average synonymous site GC content is >60% (Campos et al. 2013) and mutation is AT-biased (Figure S12), $S \rightarrow W$ SNPs should be more common overall among the introduced variants than $W \rightarrow S$ SNPs. Rapid population growth following the bottleneck would make the introduced $S \rightarrow W$ variants contribute more multiple copies of the derived *W* alleles than $W \rightarrow S$ variants, which could create the relative deficit of $W \rightarrow S$ singletons. Because this effect is expected to be stronger in regions with higher GC content, it could also explain Poh et al.'s (2012) observation that the relative deficit of $S \rightarrow W$ singletons is more apparent in highly-expressed genes.

A detailed analysis of these demographic factors is beyond the scope of this paper, as it would require knowledge of many poorly-known parameters (for example, the time and the extent of the admixture; see Duchen et al. 2013). Overall, notwithstanding the possibility that AT-ending codons may be favoured in some genes (DuMont et al. 2004; Nielsen et al. 2007), our data from a non-bottlenecked population that is close to the putative ancestor of *D. melanogaster* suggest that the genome-wide pattern is compatible with a model in which

selection on CUB is reduced in the *D. melanogaster* lineage and ongoing selection is confined to the most GC-rich parts of the genome.

In addition, Lawrie et al. (2013) suggested that a subset of 4-fold sites may be under strong selective constraints in *D. melanogaster*. These authors based their conclusions on two main observations that were made from analysing a North American population generated by the *Drosophila* Genetic Reference Panel (DGRP): a lack of difference in the shape of the SFSs between 4-fold and SI sites and a ~22% reduction in diversity level at 4-fold sites relative to SI sites (after correcting for differences in GC content; see their Figure 1). The authors suggested that their findings might represent "a largely orthogonal force to canonical codon usage bias" (p. 12 of Lawrie et al. (2013)). Indeed, by using a sample with 130 alleles, they were able to detect signals of much stronger purifying selection (with γ estimated to be -283) than is permitted by our sample sizes (21 MD lines from *D. simulans* and 17 RG lines from *D. melanogaster*). Additionally, their estimates of the intensity of strong selection appear to be uniform across genes with high and low levels of CUB, in contrast to the pattern we report here.

Obtaining more information about these two seemingly independent forces acting on 4-fold sites (weak selection on CUB and strong purifying selection) is an important area for future investigation. Several factors are of note. As discussed above, admixture is likely to complicate the analysis of the North American population of *D. melanogaster*. Although Lawrie et al. (2013) used the same method as that of Glémin et al. (2015) to control for demography, this method is nonetheless an approximation and may still lead to biased estimates of γ under certain conditions, as demonstrated by simulations (Eyre-Walker et al. 2006). Using non-admixed populations and explicit demographic models (as in this study)

may be preferable. Second, with a larger sample size (as in Lawrie et al. (2013)), it should be possible to jointly model the effects of both weak selection on CUB, which requires distinguishing $W \rightarrow S$, $S \rightarrow W$, and putatively neutral mutations (i.e., $S \rightarrow S$ and $W \rightarrow W$) (which were ignored by Lawrie et al. (2013)), and strong purifying selection, which primarily leads to an excess of very low frequency variants. By doing so, we should be able to explicitly test the relative importance of these two forces, and gain further insights into the evolution of 4-fold sites in the *Drosophila* genome.

Complex evolutionary patterns in short introns

Short introns have been widely-used as a neutral reference in *Drosophila* evolutionary genetic studies (Halligan & Keightley 2009; Parsch et al. 2010), and are thought to be closer to base composition equilibrium than other genomic regions (Kern & Begun 2005; Haddrill & Charlesworth 2008; Singh et al. 2009; Robinson et al. 2014), a pattern we have also observed (Figure 2). When analysed as a whole, the data point to the existence of a GC-favouring force in both species (Figures 3 – 4). Given the apparent lack of selective constraints in SI regions, it seems probable that gBGC may have played a significant role in their evolution. Although our detailed analyses were complicated by insufficient data, multiple aspects of the data presented in Figure 5 are nonetheless inconsistent with the standard gBGC model, which would predict that the strength of the GC-favouring effect should increase with GC content (Duret & Galtier 2009).

For *D. simulans*, the substitution patterns across short intron bins shown in Figure 5 are qualitatively similar to those shown in Figures 1 and 2 for the 4-fold sites. This seems to imply that the GC-favouring force acting on short introns may also have experienced a reduction in strength. However, in contrast to the 4-fold sites for which a genome-wide

excess of $S \rightarrow W$ substitutions was observed (Figure 2), we obtained contrasting patterns in low-GC and high-GC short intron bins (Figure 5B), with the former having a significant bias towards $W \rightarrow S$ substitutions (χ^2 test, $p = 2.30 \times 10^{-16}$), and the latter a significant bias towards $S \rightarrow W$ substitutions (χ^2 test, $p = 2.95 \times 10^{-24}$). These contrasting patterns could potentially be explained by an increase in the strength of the GC-favouring force in the low-GC short introns, but a decrease in the high-GC ones. The difference between the γ values estimated by the Glémin method and the ZC method gives some tantalising indications that this might have happened (Figure 5D). However, we are unaware of any direct evidence supporting this possibility, and it is also hard to reconcile with what we observed at the 4-fold sites, which were extracted from the same set of genes. Furthermore, the Glémin model provides little evidence that *S* basepairs are more favoured in high GC content regions, although this might have been the case in the past according to the ZC model.

In *D. melanogaster* (Figure 5F), a bias towards fixing *W* basepairs was observed in the first four SI bins (χ^2 test, maximum $p = 5.85 \times 10^{-8}$), but not the last bin (χ^2 test, $p = 0.40$). Again this is inconsistent with the genome-wide fixation bias towards *W* at the 4-fold sites (Figure 2). Estimates of γ from the two polymorphism-based methods are closer to each other compared to *D. simulans*, and both methods seem to suggest that *S* basepairs are more favoured in GC-rich regions (Figure 5H), but the small number of bins makes it difficult to draw definitive conclusions from correlation-based analyses.

To investigate this further, we calculated the polymorphism-to-divergence ratio for $W \rightarrow S$ changes, $S \rightarrow W$ changes, and changes that are supposedly unaffected by gBGC (i.e., $W \rightarrow W$ and $S \rightarrow S$ changes), denoted by $rpd_{W \rightarrow S}$, $rpd_{S \rightarrow W}$, and rpd_{neu}, respectively. If high GC content is driven by gBGC, we expect $rpd_{neu}/rpd_{W \rightarrow S} > 1$ (i.e., fixation bias

towards *S*) and $rpd_{neu}/rpd_{S\rightarrow W} < 1$ (i.e., fixation bias against *W*) in high GC bins, but these two ratios should be close to one in low GC bins where gBGC should be weak. In *D. melanogaster*, the first prediction was met ($rpd_{neu}/rpd_{W\rightarrow S}$ = 1.60, p = 0.001 and $rpd_{neu}/rpd_{S\rightarrow W}$ = 0.69, p = 7.2 x 10^{-3}, in the most GC-rich bin). However we found evidence for the existence of an AT-favouring force in the bin with the lowest GC content ($rpd_{neu}/rpd_{W\rightarrow S}$ = 0.66, p = 7.60 x 10^{-5}, and $rpd_{neu}/rpd_{S\rightarrow W}$ = 2.01, p = 3.50 x 10^{-12}), which is in agreement with estimates produced by the ZC method (Figure 5H), but inconsistent with the gBGC model. In a similar analysis of the SI bins in *D. simulans*, none of the polymorphism-to-divergence ratios were found to be significantly different from 1, except in the bin with the lowest GC content where $rpd_{neu}/rpd_{W\rightarrow S}$ = 1.25 (p = 0.0079). These findings are again inconsistent with the gBGC model.

Overall, the data from both species suggest that there is heterogeneity in evolutionary patterns between short introns residing in different parts of the genome, and that there might be some GC-favouring forces acting on short introns. However, there are substantial uncertainties as to how much of the GC-favouring effect is caused by gBGC. This conclusion is consistent with several previous studies that found little or no evidence for gBGC in *D. melanogaster* (Clemente & Vogl 2012b; Comeron et al. 2012; Campos et al. 2013; Robinson et al. 2014). Furthermore, in contrast to the 4-fold sites, where a reduction in γ is clear when estimates from the Glémin model and the ZC model are compared, no clear evidence of such a difference can be seen in the SI data. Regardless, this GC-favouring force acting on short introns is unlikely to be the sole explanation of the results obtained from 4-fold sites, because the γ estimates obtained from the latter are consistently higher than those from the former (Figure 4 vs. Figure 5). Given the importance of these putatively neutral sites in short introns, more work is necessary to understand the unique features reported above.

Acknowledgements

The authors would like to thank Peter Andolfatto, Juraj Bergman, Claus Vogl and three anonymous reviewers for helpful comments on the manuscript. Juraj Bergman and Claus Vogl also kindly provided the substitution data in Supplementary table S2. This work was supported by a PhD studentship, jointly funded by the National Environmental Research Council [grant numbers NE/H524881/1, NE/K500914/1] and the Department of Animal and Plant Sciences, University of Sheffield to BCJ. JC is supported by grant number RPG-2015-033 from the Leverhulme Trust to Brian Charlesworth. The *D. simulans* sequence data generated in the Charlesworth lab were funded by a grant from the Biotechnology and Biological Sciences Research Council to BC [grant number BB/H006028/1]. The raw reads for the newly described *D. simulans* lines are deposited in the European Nucleotide Archive, study accession number: PRJEB7673. This project made use of the computational resources provided by the University of Sheffield's high-performance computer cluster, Iceberg.

References

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

Akashi H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans* reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics. 144:1297–1307.

Akashi H. 1997. Codon bias evolution in *Drosophila*: Population genetics of mutation-selection drift. Gene. 205:269–278.

Akashi H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene. 238:39–51.

Akashi H et al. 2006. Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics. 172:1711–26.

Akashi H, Goel P, John A. 2007. Ancestral inference and the study of codon bias evolution: implications for molecular evolutionary analyses of the *Drosophila melanogaster* subgroup. PLoS One. 2:e1065.

Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of silent DNA polymorphism in *Drosophila*. Genetics. 146:295–307.

Andolfatto P, Wong KM, Bachtrog D. 2011. Effective population size and the efficacy of selection on the X chromosomes of two closely related *Drosophila* species. Genome Biol. Evol. 3:114–28.

Begun DJ et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLOS Biol. 5:e310.

Bergland AO, Tobler R, González J, Schmidt P, Petrov D. 2016. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. Mol. Ecol. 25:1157–74.

Bierne N, Eyre-Walker A. 2006. Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome. J. Evol. Biol. 19:1–11.

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

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. Genetics. 159:1779–1788.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. 31:1010–28.

Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2013. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster*. Mol. Biol. Evol. 30:811–23.

Caracristi G, Schlötterer C. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles.

Mol. Biol. Evol. 20:792–9.

Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77:153–166.

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.

Charlesworth B. 2013. Stabilizing selection, purifying selection, and mutational bias in finite populations. Genetics. 194:955–71.

Charlesworth B, Charlesworth D. 2010. *Elements of evolutionary genetics*. Roberts and Company Publishers: Greenwood Village.

Clemente F, Vogl C. 2012a. Evidence for complex selection on four-fold degenerate sites in *Drosophila melanogaster*. J. Evol. Biol. 25:2582–95.

Clemente F, Vogl C. 2012b. Unconstrained evolution in short introns? - an analysis of genome-wide polymorphism and divergence data from *Drosophila*. J. Evol. Biol. 25:1975–90.

Collins TM, Wimberger PH, Naylor GJP. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. 43:482–496.

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

Comeron JM, Williford A, Kliman RM. 2008. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity. 100:19–31.

Cutter AD, Charlesworth B. 2006. Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. Curr. Biol. 16:2053–7.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14:262–274.

Dean MD, Ballard JWO. 2004. Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. Mol. Phylogenet. Evol. 32:998–1009.

DePristo MA et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 134:341–52.

Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S. 2013. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics. 193:291–301.

DuMont VB, Fay J, Calabrese P, Aquadro C. 2004. DNA variability and divergence at the Notch locus in *Drosophila melanogaster* and *D. simulans*: A case of accelerated synonymous site divergence. Genetics. 167:171–185.

Dunn KA, Bielawski JP, Yang Z. 2001. Substitution rates in *Drosophila* nuclear genes: Implications for translational selection. Genetics. 157:295–305.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10:285–311.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. 96:4482–4487.

Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, Melnick DJ. 2014. Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys. Mol. Biol. Evol. 31:2425–40.

Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. J. Mol. Evol. 47:686–690.

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

Fuller ZLZ et al. 2014. Evidence for stabilizing selection on codon usage in chromosomal rearrangements of *Drosophila pseudoobscura*. G3. 4:2433–49.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. Genetics. 172:221–8.

Glémin S et al. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res. 25:1215–1228.

Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. Biol. Lett. 4:438–41.

Haddrill PR, Zeng K, Charlesworth B. 2011. Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. Mol. Biol. Evol. 28:1731–43.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res. 16:875–84.

Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. 40:151–172.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol. Biol. Evol. 24:2196–202.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu. Rev. Genet. 42:287–99.

Hey J, Kliman R. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics. 160:595–608.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89–98.

Jackson BC, Campos JL, Zeng K. 2015. The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. Heredity. 114:163–74.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Keightley PD et al. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195–201.

Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol. Biol. Evol. 22:51–62.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press: Cambridge, UK.

Kliman RM. 1999. Recent selection on synonymous codon usage in *Drosophila*. J. Mol. Evol. 49:343–351.

Kliman RM. 2014. Evidence that natural selection on codon usage in *Drosophila pseudoobscura* varies across codons. G3. 4:681–92.

Kofler R et al. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 6:e15925.

Langley CH et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics. 192:533–98.

Langley SA, Karpen GH, Langley CH. 2014. Nucleosomes shape DNA polymorphism and divergence. PLoS Genet. 10:e1004457.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet. 9:e1003527.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–9.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–60.

Li W-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24:337–345.

Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. 5:R45.

Maside X, Lee AW, Charlesworth B. 2004. Selection on Codon Usage in *Drosophila americana*. Curr. Biol. 14:150–154.

Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. Genetics. 200:873–90.

Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. 2016. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Mol. Biol. Evol. 33:1580–9.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 351:652–654.

McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74:145–158.

McVean GAT, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics. 155:929–944.

McVean GAT, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics. 157:245–257.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. U. S. A. 110:8615–20.

Nielsen R et al. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol. Biol. Evol. 24:228–35.

Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics. 26:419–20.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. Mol. Biol. Evol. 27:1226–34.

Poh Y-P, Ting C-T, Fu H-W, Langley CH, Begun DJ. 2012. Population genomic analysis of base composition evolution in *Drosophila melanogaster*. Genome Biol. Evol. 4:1245–55.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution. 61:3001–6.

Pool JE, Nielsen R. 2008. The impact of founder events on chromosomal variability in multiply mating species. Mol. Biol. Evol. 25:1728–36.

Powell J, Moriyama E. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. 94:7784–7790.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. 13:735–748.

Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. Mol. Biol. Evol. 31:425–33.

Rogers RL et al. 2014. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. Mol. Biol. Evol. 31:1750–66.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics. 132:1161–1176.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics. 194:937–54.

Shields D, Sharp P, Higgins D, Wright F. 1988. 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5:704–716.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. Mol. Biol. Evol. 26:1591–605.

Singh ND, Arndt PF, Petrov DA. 2005. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics. 169:709–22.

Singh ND, Macpherson JM, Jensen JD, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. BMC Evol. Biol. 7:202.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol. Biol. Evol. 28:63–70.

Stone EA. 2012. Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. Genome Res. 22:966–74.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–95.

Takano-Shimizu T. 1999. Local recombination and mutation effects on molecular evolution in *Drosophila*. Genetics. 153:1285–1296.

Takano-Shimizu T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* Chromosomes. Mol. Biol. Evol. 18:606–619.

True JR, Mercer JM, Laurie CC. 1996. Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics. 142:507–523.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol. Biol. 7:226.

Vogl C, Bergman J. 2015. Inference of directional selection and mutation parameters assuming equilibrium. Theor. Popul. Biol. 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. Theor. Popul. Biol. 81:197–209.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–91.

Yukilevich R, Turner TL, Aoki F, Nuzhdin S V, True JR. 2010. Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster.* Genetics. 186:219–39.

Zeng K. 2010. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. Mol. Biol. Evol. 27:1327–37.

Zeng K. 2012. The application of population genetics in the study of codon usage bias. In: Codon evolution: mechanisms and models. Cannarozzi, GM & Schneider, A, editors. Oxford University Press pp. 245–254.

Zeng K, Charlesworth B. 2009. Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics. 183:651–662.

Zeng K, Charlesworth B. 2010a. Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. J. Mol. Evol. 70:116–128.

Zeng K, Charlesworth B. 2010b. The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics. 186:1411–24.

Zhang C, Wang J, Long M, Fan C. 2013. gKaKs: the pipeline for genome-level Ka/Ks calculation. Bioinformatics. 29:645–6.

Table 1. Summary statistics for the filtered *D. simulans* dataset.

All statistics were calculated per gene, and the means are presented here.

Chr.[a]	Site	Within-population statistics					Population differentiation
		Pop.[b]	π[c]	θ_W[d]	Δ_π[e]	D[f]	F_{ST}
A	0-fold[g]	MD	0.0016	0.00269	-0.12	-1.29	0.0202
		NS	0.00148	0.00206	-0.0882	-0.903	
	4-fold[h]	MD	0.0317	0.0434	-0.0784	-1.03	0.0252
		NS	0.0294	0.0347	-0.0457	-0.579	
	SI[i]	MD	0.0321	0.0417	-0.065	-0.603	0.0174
		NS	0.0297	0.0340	-0.036	-0.326	
X	0-fold	MD	0.00119	0.00207	-0.125	-1.27	0.0178
		NS	0.00113	0.00163	-0.0942	-0.924	
	4-fold	MD	0.0182	0.0282	-0.104	-1.31	0.0246
		NS	0.0173	0.0225	-0.0706	-0.847	
	SI	MD	0.0208	0.0298	-0.0924	-0.785	0.0194
		NS	0.0195	0.0248	-0.0591	-0.509	

[a] Chromosome

[b] Population sample: MD – Madagascar; NS – Kenya

[c] Average number of pairwise differences between lines

[d] Watterson's estimator of θ, the scaled mutation rate

[e] See equation (1)

[f] Tajima's D

[g] 0-fold degenerate sites

[h] 4-fold degenerate sites

[i] Sites 8-30bp of introns < 66bp in length

Table 2. Counts of substitutions along the *D. melanogaster* and *D. simulans* lineages at 4-fold degenerate and SI sites.

		D. simulans				*D. melanogaster*			
		A		X		A		X	
Site[a]	Polarisation method[b]	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT
4-fold	parsimony	13607	25656	1962	3934	10588	40586	1140	7395
	SBR	14085	30524	2116	4528	11285	47894	1258	8670
	AWP	15219	30945	2450	4639	12399	48264	1425	8611
SI	parsimony	1859	1598	206	152	1570	1884	131	229
	SBR	1930	2052	231	183	1658	2417	146	271
	AWP	2006	2158	217	206	1718	2506	141	303

[a] 4-fold – 4-fold degenerate sites; SI – Sites 8-30bp of introns < 66bp in length

[b] The ancestral state at the *melanogaster-simulans* node was determined using three methods: parsimony, the single best reconstruction (SBR) under the GTR-NH_b model implemented in PAML, and the average weighted by posterior probability (AWP) under the GTR-NH_b model implemented in PAML

Figure Legends

Fig 1. **Substitution rates.** The results are shown for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points), binned according to the GC content of the extant *D. melanogaster* reference sequence. Rates were calculated for the *D. simulans* lineage (top row) and the *D. melanogaster* lineage (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column). AT → GC substitutions – teal circles; GC → AT substitutions – orange triangles.

Fig 2. The ratios of substitution counts. The results are shown for positions 8-30bp of introns <66bp long (SI sites; leftmost point), and 4-fold degenerate sites (remaining points), binned as described for Fig 1. A substitution count ratio of $N_{W\to S}/N_{S\to W} = 1$ implies equilibrium base composition. Ratios were calculated for the *D. simulans* lineage (top row) and the *D. melanogaster* lineage (bottom row), for autosomes (left-hand column) and X (right-hand column).

Fig 3. **Derived allele frequencies**. Mean DAFs are shown for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points), binned as described for Fig 1. Mean DAFs were calculated using the MD (Madagascan) sample of *D. simulans* (top row) and the RG (Rwandan) sample of *D. melanogaster* (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column). AT → GC mutations – teal circles; GC → AT mutations – orange triangles; AT → AT mutations or GC → GC mutations – lilac squares.

Fig 4. **The estimated strength of selection favouring GC alleles.** The estimates of the strength of selection in favour of GC alleles ($\gamma = 4N_e s$) are shown for positions 8-30bp of introns <66bp long (SI sites; leftmost points), and 4-fold degenerate sites (remaining points), binned as described for Fig 1. γ was estimated using the MD (Madagascan) sample of *D. simulans* (top row) and the RG (Rwandan) sample of *D. melanogaster* (bottom row), for autosomes (left-hand column) and X-linked sites (right-hand column). Two methods were used: the method of Zeng and Charlesworth (2009) with a one-step size in population size (ZC in the main text) – green circles; and the method of Glémin et al. (2015), not

incorporating polarisation errors (M1 in the main text) – pink squares. Filled points – bins where a model with $\gamma \neq 0$ fitted best; open points – bins where a model with $\gamma = 0$ fitted best.

Figure 5. Results for autosomal short intronic (SI) sites binned by GC content. Top row – data from the MD (Madagascan) sample of *D. simulans*; bottom row – data from the RG (Rwandan) sample of *D. melanogaster*. Panels A and E – substitution rates for AT → GC substitutions (teal circles) and GC → AT substitutions (orange triangles). Panels B and F – the ratio of substitution counts along each lineage. Panels C and G – derived allele frequencies (DAF) for AT → GC mutations (teal circles); GC → AT mutations (orange triangles); AT → AT mutations or GC → GC mutations (lilac squares). AT → AT and GC → GC mutations were labelled as neutral to signify that they should be unaffected by gBGC. Panels D and H – estimated values of the magnitude of selection in favour of GC alleles ($\gamma = 4N_e s$). Two methods were used: the method of Zeng and Charlesworth (2009) with a one-step size in population size (ZC in the main text) – green circles; and the method of Glémin et al. (2015), not incorporating polarisation errors (M1 in the main text) – pink squares. Filled points – bins where a model with $\gamma \neq 0$ fitted best; open points – bins where a model with $\gamma = 0$ fitted best. All analyses that required reconstruction of the ancestral state at the *ms* node used the AWP method, as described in the main text.

Figure 1.

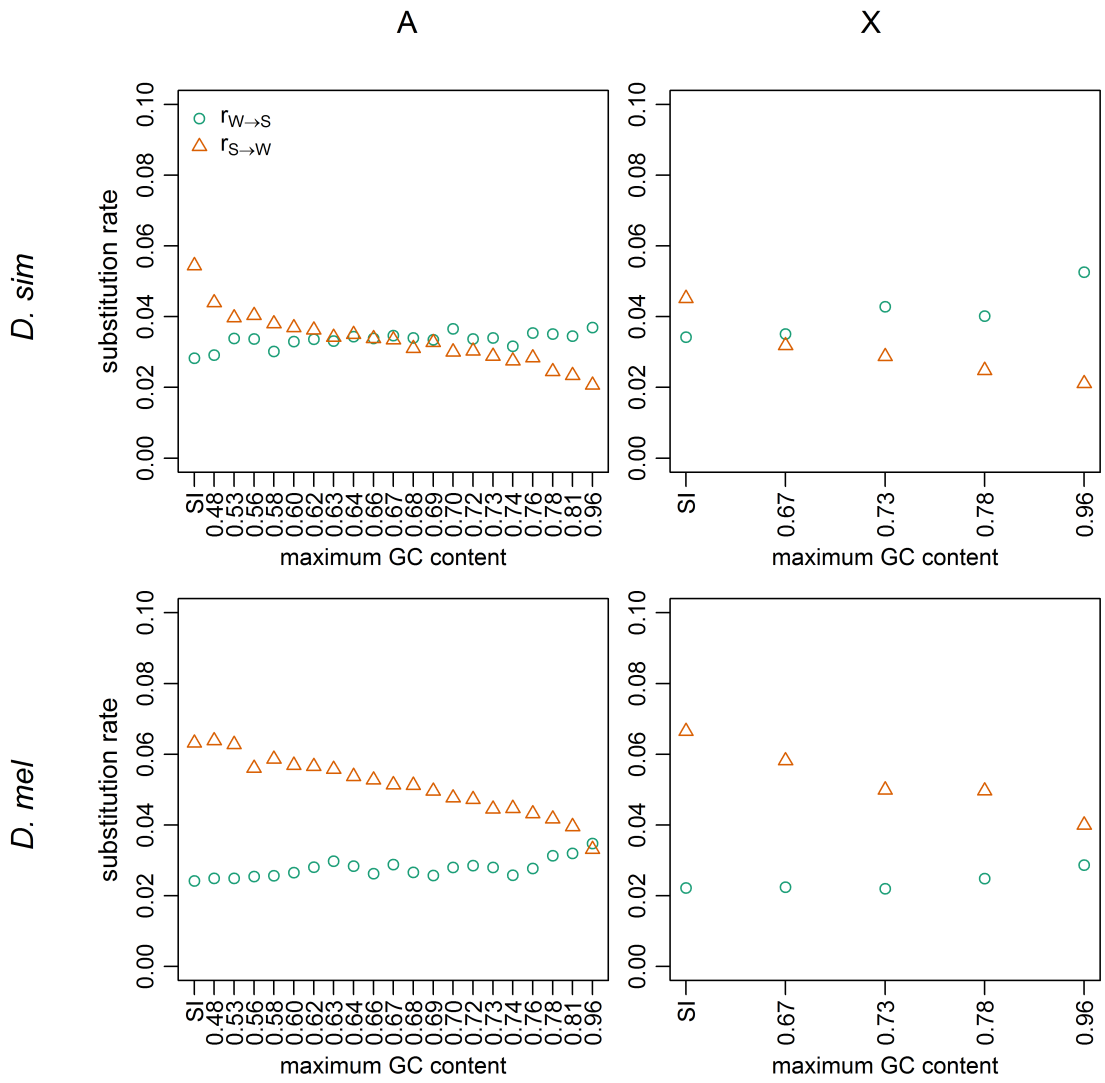

A
X
D. sim
D. mel
substitution rate
0.00
0.02
0.04
0.06
0.08
0.10
$r_{W\to S}$
$r_{S\to W}$
SI
0.48
0.53
0.56
0.58
0.60
0.62
0.63
0.64
0.66
0.67
0.68
0.69
0.70
0.72
0.73
0.74
0.76
0.78
0.81
0.96
0.67
0.73
0.78
0.96
maximum GC content

Figure 2.

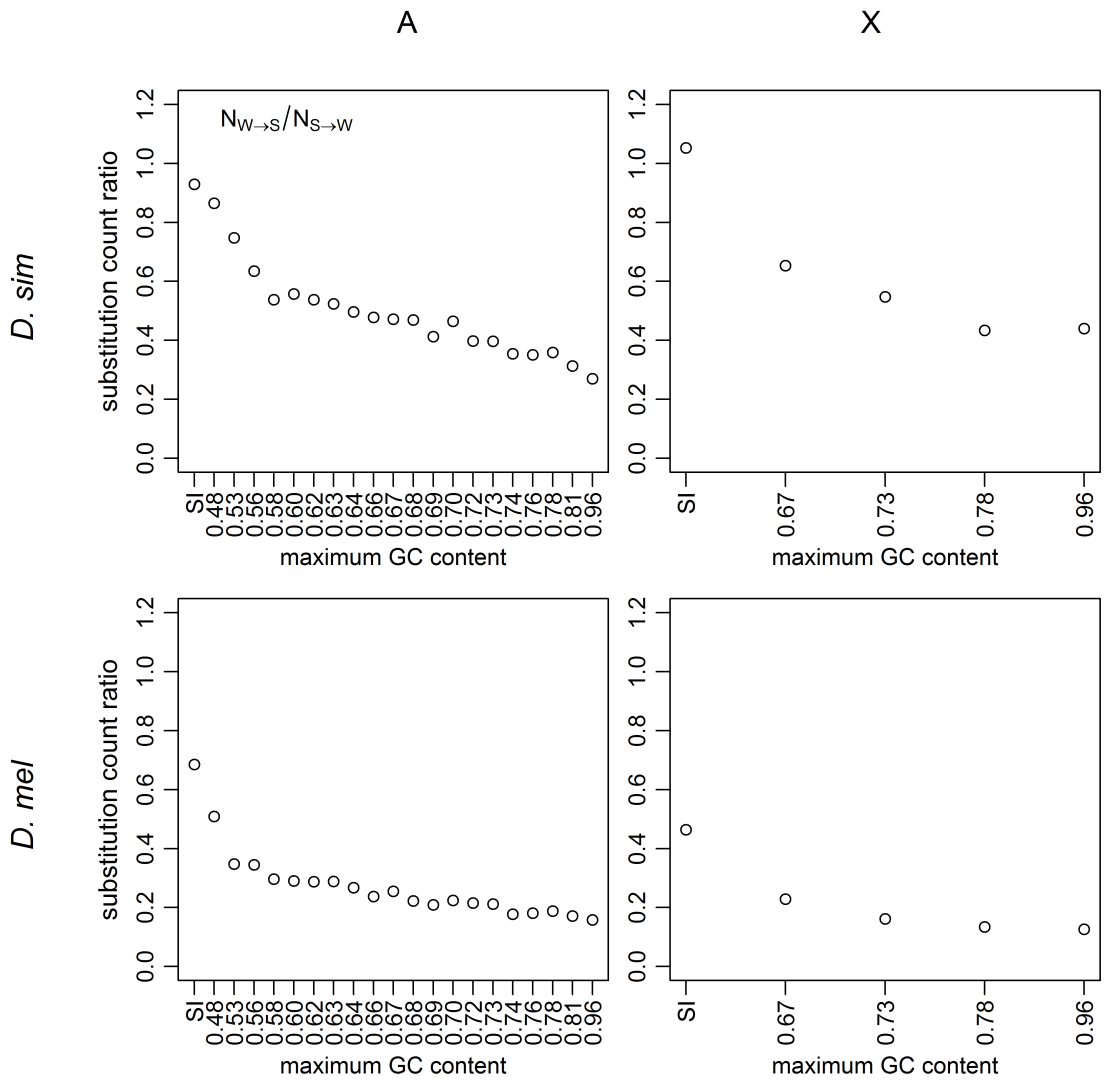

A
X
D. sim
D. mel
substitution count ratio
$N_{W\to S}/N_{S\to W}$
maximum GC content
SI
0.48
0.53
0.56
0.58
0.60
0.62
0.63
0.64
0.66
0.67
0.68
0.69
0.70
0.72
0.73
0.74
0.76
0.78
0.81
0.96
0.67
0.73
0.78
0.96

Figure 3.

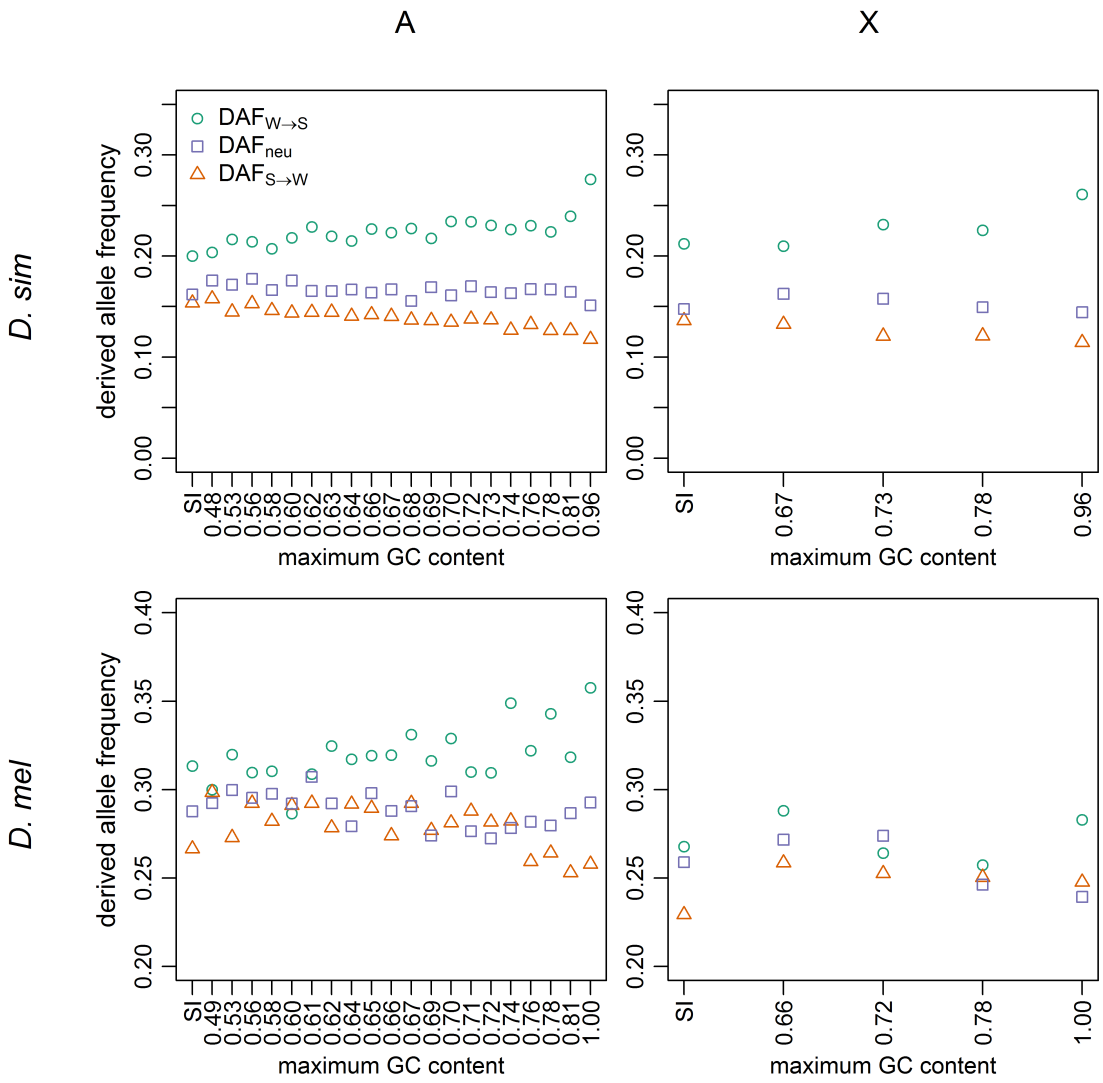

A
X
D. sim
D. mel
DAF$_{W\rightarrow S}$
DAF$_{neu}$
DAF$_{S\rightarrow W}$
derived allele frequency
maximum GC content
SI
0.48
0.53
0.56
0.58
0.60
0.62
0.63
0.64
0.66
0.67
0.68
0.69
0.70
0.72
0.73
0.74
0.76
0.78
0.81
0.96
0.67
0.73
0.78
0.96
0.49
0.61
0.65
0.71
1.00
0.66
0.72

Figure 4.

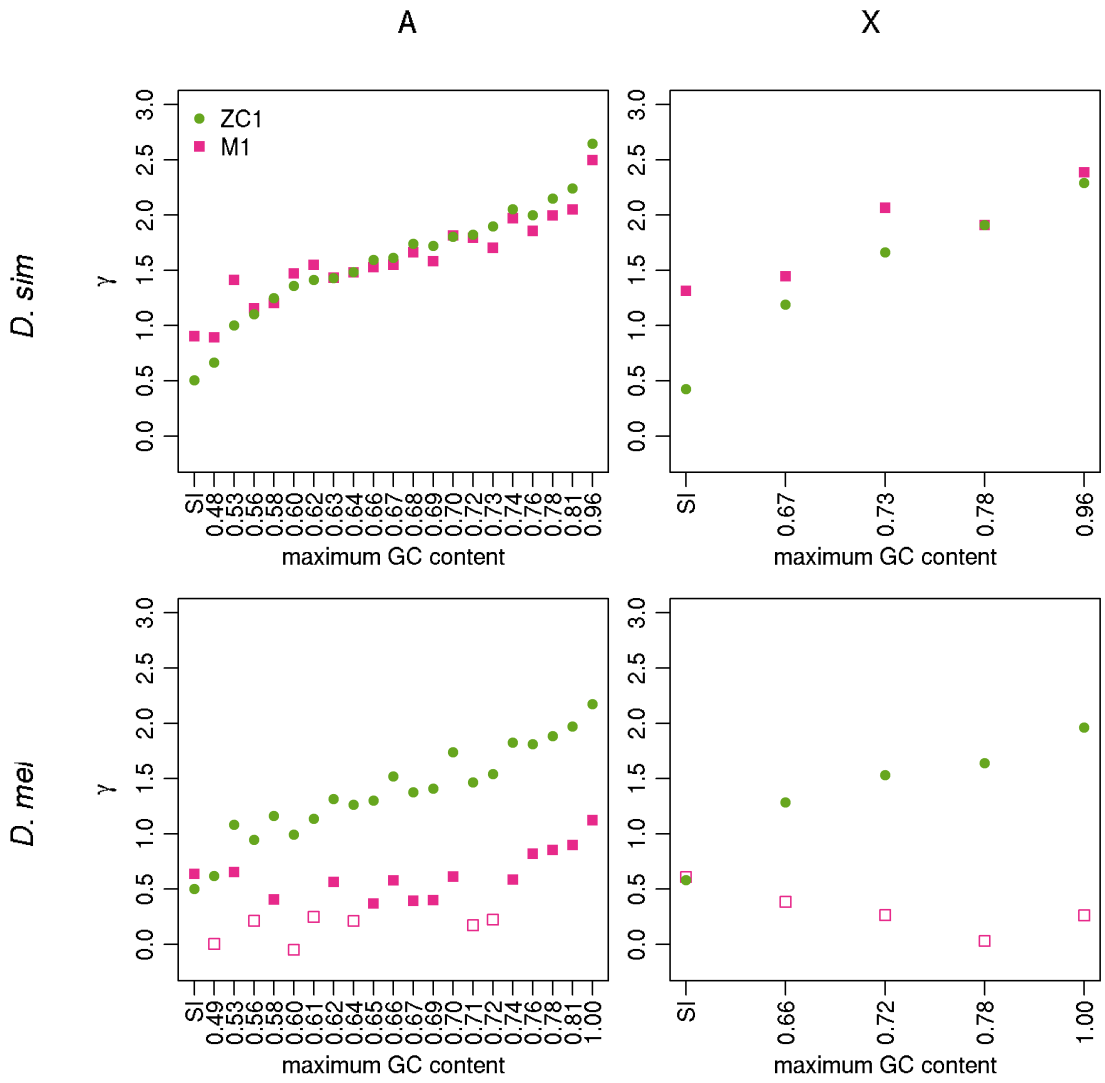
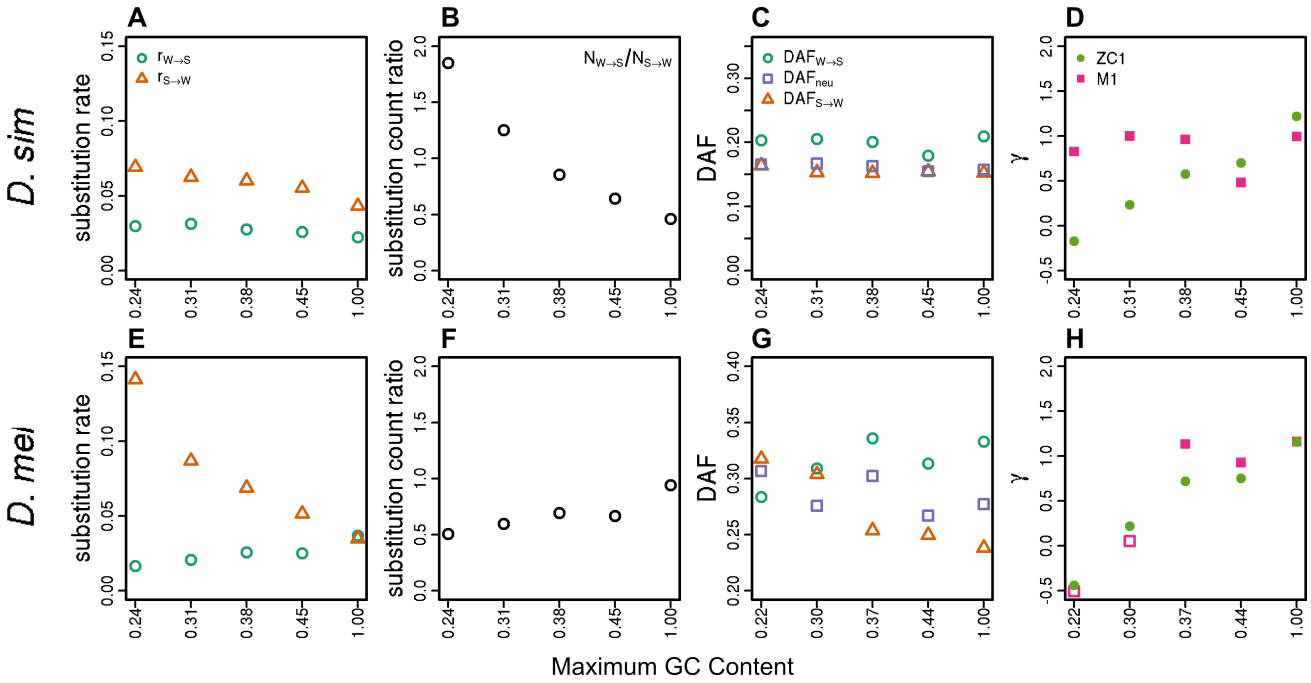

A
X
D. sim
D. mel
ZC1
M1
γ
maximum GC content
SI
0.48
0.53
0.56
0.58
0.60
0.62
0.63
0.64
0.66
0.67
0.68
0.69
0.70
0.72
0.73
0.74
0.76
0.78
0.81
0.96
0.67
0.73
0.78
0.96
0.49
0.61
0.65
0.71
1.00
0.66
0.72

Figure 5.


A
B
C
D
E
F
G
H
D. sim
D. mel
substitution rate
substitution count ratio
DAF
γ
$r_{W\to S}$
$r_{S\to W}$
$N_{W\to S}/N_{S\to W}$
$DAF_{W\to S}$
DAF_{neu}
$DAF_{S\to W}$
ZC1
M1
Maximum GC Content