GC-biased segregation of non-coding polymorphisms in Drosophila.

Nicolas Galtier, Eric Bazin, and Nicolas Bierne.

UMR 5171 – "Génome, Populations, Interactions, Adaptation"

CNRS - Université Montpellier 2 - IFREMER

34095 Montpellier, France
Running head: GC-bias in Drosophila non-coding SNPs

Key Words: GC-content, biased gene conversion, sequence polymorphism, non-coding DNA, codon usage

Corresponding author:

N. Galtier
CNRS UMR 5171 – "Génome, Populations, Interactions, Adaptation"
Université Montpellier 2 – CC 63
Place E. Bataillon
34095 Montpellier, France

phone: (+33) 467 14 46 84
fax: (+33) 467 14 45 54
email: galtier@univ-montp2.fr
Abstract

The study of base composition evolution in Drosophila has been mostly achieved through the analysis of coding sequences. Third codon position GC-content, however, is influenced by both neutral forces (e.g. mutation bias) and natural selection for codon usage optimisation. In this paper, large data sets of non-coding DNA sequence polymorphism in *D. melanogaster* and *D. simulans* were gathered from public data bases to try and disentangle these two factors – non-coding sequences are not affected by selection for codon usage. Allele frequency analyses revealed an asymmetric pattern of AT vs. GC non-coding polymorphisms: AT→GC mutations are less numerous, and tend to segregate at a higher frequency, than GC→AT ones, especially at GC-rich loci. This is indicative of non-stationary evolution of base composition and/or GC-biased allele transmission. Fitting population genetics models to the allele frequency spectra confirmed this result, and favoured the hypothesis of a biased transmission. These results, together with previous reports, suggest that GC-biased gene conversion has influenced base composition evolution in Drosophila, and explains the correlation between intron and exon GC-content.
INTRODUCTION

The evolution of base composition, the percentage of A, C, G and T of genomic sequences, has been a topic of interest in a number of taxonomic groups (e.g. Jermiin et al. 1994, Bernardi 2000, Galtier & Lobry 1997, Eyre-Walker & Hurst 2001), possibly because it is a typical instance of the neutralist/selectionist debate of molecular evolution: we want to identify the evolutionary forces (e.g. mutation, natural selection, recombination) shaping base composition variation within and between genomes, and their relative importance.

In Drosophila, this question is closely linked to the issue of synonymous codon usage bias. The various codons encoding a given amino-acid are not used randomly: some codons are "preferred" over synonymous alternatives, probably because they allow more efficient / accurate translation (Akashi et al. 1998). For an unknown reason, 19 preferred codons out of 20 are ending in G or C in Drosophila (Marais et al. 2001). Understanding the evolutionary dynamics of codon usage bias, therefore, requires disentangling selection on codon choice from forces acting on the GC-content of genomic sequences irrespective of their coding/non-coding status (Marais et al. 2001, Hey & Kliman 2002, Marais et al. 2003).

The GC-content of non-coding DNA varies between regions of the Drosophila genome, and the GC-content of introns (GCi) is correlated with the GC-content at third codon position of exons (GC3; Kliman & Eyre-Walker 1998) of a gene. This correlation was correctly taken into account by Kliman and Hey (1993) when analysing codon usage data, and interpreted as the consequence of a variable mutation bias: the ratio of AT→GC over GC→AT mutation rates would vary along the genome. An alternative hypothesis was recently proposed, however: the GC3/GCi correlation could be due to GC-biased gene conversion (BGC, Marais 2003). Allelic gene conversion is a molecular process associated with recombination in which a fragment of one of the two recombining chromosomes is "copied/pasted" onto the other one – a unidirectional genetic exchange. It is a fundamental
process of DNA metabolism occurring during the repair of double-strand breaks. Although little empirical molecular evidence has been reported to date, numerous arguments indicate that gene conversion is biased toward GC in yeast (Birdsell 2002) and mammals (Galtier et al. 2001, Galtier 2003, Kudla et al. 2004, Meunier & Duret 2004): G or C alleles tend to convert A or T more frequently than the reverse, resulting in a higher fixation probability for G and C alleles. GC-content is positively, but very weakly (R²<1%), correlated to local recombination rate in *Drosophila melanogaster* (Marais et al. 2001), suggesting that BGC might also impact GC-content in this species.

Demonstrating the existence of BGC in Drosophila would be of interest for several reasons. First, this would be an additional species in which this up-to-now neglected evolutionary force applies, adding some credit to the "universality" of BGC advocated by Birdsell (2002). Secondly, this would provide a plausible explanation for the GC3/GCi correlation in Drosophila genome. Thirdly, this would ask for a reappraisal of coding-sequence polymorphism patterns in Drosophila, and of hypothesis about the relationship between selection on codon usage, GC-content and recombination.

BGC is a neutral process that mimics natural selection by conferring a higher fixation probability to G and C alleles - heterozygotes produce a larger amount of G and C than A and T gametes. This effect is not distinguishable from a selective advantage of GC over AT alleles (Nagylaky 1983). Such a selective pressure in favour of GC actually occurs at synonymous sites for codon usage optimisation. Our strategy, therefore, was to analyse polymorphism patterns in non-coding sequences of *D. melanogaster* and *D. simulans*, where selection for codon usage does not apply. In the absence of BGC and of selection for genomic GC-content – our null hypothesis – AT and GC polymorphisms should have similar allele frequency spectra.
DATA

Two non-coding sequence polymorphism data sets from *D. melanogaster* were obtained from the EMBL database (version 73, March 2004). The first data set, hereafter called "Dme_exhaustive", was built as a part of the Polymorphix database (Bazin *et al.* 2005). Nuclear sequences from *D. melanogaster* were extracted from EMBL and grouped into 283 families, first according to sequence similarity (at least 80% similarity required), then using a bibliographic criterion (a sequence S was kept in its family F only if at least 3 other sequences in F were published in the same article as S). Sequence families were inspected by eye. Those not corresponding to polymorphism studies (*e.g.* transposons) were manually deleted. Four families including paralogous loci were cleaned or split to ensure orthology. Families including less than ten sequences were removed. Sequences were aligned using the MABIOS algorithm (Abeddaim *et al.* 1997), especially efficient for a quick alignment of highly similar sequences. Alignments were slightly modified manually. Ambiguously aligned regions (*e.g.* microsatellites) were manually discarded, as well as sequences containing too many gaps or missing nucleotides. Coding regions were removed from the alignments. The "Dme_exhaustive" data set finally included 221 families (*i.e.*, loci) including 10 to 100 sequences.

The "Dme_exhaustive" data set is highly heterogeneous with respect to population sampling - worldwide for some loci, local for others. Allele frequencies are mostly unknown: what we have are haplotypes (EMBL entries), some of which being possibly shared by several individuals in a sample. This data set is therefore not suitable for standard population genetics analyses. It should be noted, however, that sampling peculiarities should affect AT→GC mutations and GC→AT mutations in a similar way. Under the null, no-BGC hypothesis, an AT/GC symmetric pattern is expected whatever the sampling, and whatever the population history.
The second *D. melanogaster* data set, hereafter called "Dme_Glinka", is a subset of the "Dme_exhaustive" one corresponding to data from Glinka *et al.* (2003). In this study, 105 X-linked DNA fragments were sequenced in 19-24 *D. melanogaster* lines from Europe and Africa. We decided to make use of African data only since natural selection significantly affects polymorphism patterns in the European data set (Glinka *et al.* 2003). The "Dme_Glinka" data set is restricted to X-linked loci, and to a single African population, but its homogeneous sampling allows more elaborate analyses using population genetics models. Model fitting was performed on a subset of 94 loci for which exactly 12 African lines have been sampled and sequenced.

A similar strategy was conducted in *D. simulans*. A "Dsi_exhaustive" data set of 58 loci was built by gathering sequences from Polymorphix. A minimal number of eight sequences per family were required. Again, the "Dsi_exhaustive" data set is too heterogeneous to allow proper population genetics model fitting. A data subset called "Dsi_Begun" was built by using data obtained by D. Begun's group only (41 loci). These studies used samples from a single Californian population of *D. simulans*. All these data sets are available from http://kimura.univ-montp2.fr/data.

**ANALYSIS**

*D. melanogaster*: The "Dme_exhaustive" data set included 4672 polymorphic sites, among which 3364 A or T vs. G or C biallelic sites. Figure 1 shows the distribution of the observed frequency of the G or C state in these sites. Under the hypothesis of unbiased transmission (i.e., no BGC, no selection) and stationary GC-content, this distribution should be symmetric (Eyre-Walker 1999, Lercher *et al.* 2002). This was not the case here. The number of sites for which the G or C state occurs at a frequency strictly higher than 0.5 – call them GC-sites -
was 1875, whereas 1408 sites showed a frequency of G or C strictly lower than 0.5. The difference is highly significant (p<10^{-8}, binomial test).

Then loci from the "Dme_exhaustive" data set were split in an arbitrary three categories of equal size according to their GC-content, and the above binomial test was re-conducted. A significant excess of GC-sites was detected in the high-GC (expected 501, observed 626, p<10^{-10}) and medium-GC (expected 672, observed 765, p<10^{-6}) data subsets, but not in the low-GC data subset (expected 468, observed 484), indicating that the evolutionary force causing this asymmetric pattern is stronger in GC-rich regions. This trend was confirmed when we plotted the difference between the numbers of AT-and GC-sites of a locus (normalized by the total amount of polymorphism) vs. locus GC-content (figure 2, \( r^2=0.167, p<10^{-4} \)). GC-rich loci showed an excess of AT-mutations, while AT-rich loci appeared to be at equilibrium.

The detected GC-biased segregation of polymorphisms must be caused by a departure from at least one of the two assumptions of the null model, namely unbiased transmission and stationary GC-content. Transmission distortion in favour of GC would lead to an increased number of GC-sites (Eyre-Walker 1999), consistent with the asymmetric figure 1. Alternatively, the observed pattern could be explained in a neutral context by an excess of GC→AT over AT→GC mutations, i.e., a decrease of the genomic GC-content in the GC-richest regions. Both hypothesis predict a higher than expected number of GC-sites, but make distinct predictions about the shape of the allele frequency distribution, as illustrated by figure 3. We made use of the "Dme_Glinka" data set to try and distinguish between the two scenarios.

Loci in the "Dme_Glinka" data set were split in three GC-content categories, as explained above. The total numbers of AT vs. GC segregating sites were 524, 505 and 372 for the low-GC (average GC-content: 33.5%), medium-GC (39.4%) and high-GC (47.8%)
categories, respectively. The distributions of G or C allele frequency in the three data subsets are shown (figure 4). Just like for the "Dme_exhaustive" data set, the AT/GC asymmetry appears stronger for GC-rich than GC-poor loci.

Four population genetics models were fit to each of these data subsets. The most general model is M3, in which we simply assume that sites are independent from each other, and that the individuals were sampled from a panmictic population at mutation/transmission distortion/drift equilibrium, possibly with non-stationary GC-content (where transmission distortion refers either to natural selection or to BGC). This model is described by two parameters, namely $W = 3N_e w$, the population transmission distortion coefficient in favour of C and G alleles in the X chromosome, and $\alpha = \mu_{GC}/(\mu_{AT} + \mu_{GC})$, the probability that a GC$\leftrightarrow$AT mutation is an AT$\rightarrow$GC one ($\mu_{GC}$ is the mutation rate from AT to GC, and $\mu_{AT}$ the mutation rate from GC to AT).

For a given $W$ and $\alpha$, the proportion of polymorphic sites consisting of $i$ C or G and $12-i$ A or T ($0 < i < 12$) expected under M3, $f^*(i)$, was calculated by adapting the theory in Lercher et al. (2002). Equations are given in Appendix. The log-likelihood of the model was computed using the multinomial formula:

$$\ln L = C + \sum_{i=1}^{11} f(i) \cdot \ln(f^*(i))$$

(1)

where $C$ is a constant, and where $f(i)$ is the observed number of sites consisting of $i$ C or G and $12-i$ A or T. This method is very similar to the approach of Lercher et al. (2002) and Duret et al. (2002).

The other models investigated are special instances of model M3. Model M2 (one parameter) assumes stationarity, which is obtained by forcing $\alpha$ to be such that the
equilibrium GC-content, given $W$, equals the observed GC-content. Model $M_1$ (one parameter) allows for non-stationary evolution, but assumes neutrality (no transmission distortion). This is obtained by leaving $\alpha$ free but setting $S=0$. Model $M_0$ (zero parameter), finally, assumes both stationarity and neutrality. The likelihood under models $M_0$, $M_1$ and $M_2$ was calculated just like under $M_3$.

Parameters were estimated by the maximum likelihood method. The likelihood curve (or surface) was traced using a grid on the parameter space. To maintain parameter values within a reasonable range, the maximisation was performed under the constraint that the equilibrium GC-content had to be within 27% and 60%, the lowest and highest GC-content of the analysed loci. This was necessary for model $M_3$, for which the likelihood was maximal for a very high $W$ and a very low $\alpha$, implying an irrelevant equilibrium GC-content close to 100%. For each model, the log-likelihood for the whole data set was obtained by summing the log-likelihoods of the three data subsets. Models were compared using likelihood ratio tests: if $M$ ($n$ parameters) is the null model, and $M'$ ($n'$ parameters, $n'>n$) an alternative model such that $M$ is a special case of (nested in) $M'$, then twice the difference of log-likelihood between the two models is asymptotically distributed as a $\chi^2$ with $n'-n$ degrees of freedom. The 1% significance level was retained.

The results are given in table 1. When the three data subsets were considered jointly, models $M_1$ (non-stationary) and $M_2$ (transmission distortion) vastly improved the fit over the null model $M_0$ ($M_2$ vs. $M_0$: $2.\Delta \ln L=71.2$, 3 df, $p<10^{-6}$), confirming the asymmetric nature of the distributions in figure 4. Model $M_2$ fit the data better than model $M_1$, but the significance of this difference could not be directly assessed by a likelihood ratio test since the two models are not nested. Model $M_3$ (non-stationary + transmission distortion) did improve the fit over $M_1$ ($2.\Delta \ln L=20.3$, 3 df, $p=1.5 \times 10^{-4}$), but only marginally over $M_2$ ($\Delta \ln L=9.9$, 3 df, $p=1.9 \times 10^{-2}$). $M_2$ was the model favoured by Akaike's criterion of model choice (not shown). These results
indicate that the AT/GC asymmetry of SNP segregation cannot be explained by a departure from the stationarity hypothesis only. Transmission distortion, when invoked, significantly improved the fit as compared to the neutral model, either under stationarity (M2 vs. M0) or under non-stationarity (M3 vs. M1).

Similar results were found when the three GC-categories were considered separately. The medium-GC and high-GC categories revealed a strong asymmetry, best explained by transmission distortion. No strong departure from the neutral, stationary M0 model was detected from the low-GC data set, however, in agreement with results from the "Dme_exhaustive" data set. The population transmission distortion coefficient estimated under M2 increased from 0.3 (low-GC data subset) to 1.2 (high-GC data subset).

A goodness-of-fit test was conducted by fitting a degenerate model in which every allele frequency class has its own free-to-vary frequency (11 parameters). This model significantly improved the fit over any of the M0 to M3 models (not shown). This is indicative of a poor fit of the population genetics models to the "Dme_Glinka" data set.

**D. simulans:** The "Dsi_exhaustive" data set yielded results comparable to the "Dme_exhaustive" one (figure 1): the number of GC-sites (354) was significantly higher ($p = 5.1 \times 10^{-4}$) than the number of AT-sites (271). The proportion of GC-sites (66.6%) was nearly identical to that obtained from the "Dme_exhaustive" data set (67.0%). When the 58 loci were split into 29 GC-poor and 29 GC-rich ones, only the GC-rich subset yielded a significant excess of GC-sites (209 observed, 174 expected, $p = 1.3 \times 10^{-4}$), while GC-poor loci appeared to be at equilibrium (145 observed, 138 expected).

We then attempted to fit models M0 to M3 to the "Dsi_Begun" data set. Neither when all 41 loci were analysed jointly nor when they were split in a GC-poor and a GC-rich subset did we detect any significant departure from the null hypothesis of neutrally evolving,
equilibrium GC-content ($M_0$), in contrast with the "Dme_Glinka" analysis. There are, however, two important differences between the two data sets: the "Dme_Glinka" data set is larger (1249 polymorphic sites, vs. 380 in "Dsi_Begun"), and includes a higher proportion of GC-rich loci (33% of loci have a GC-content higher than 42%, while this proportion is 19% in "Dsi_Begun"). When the eight loci from "Dsi_Begun" having a GC-content higher than 42% were analysed, a (marginally) significant departure from the null hypothesis was detected, and alternative models were favoured ($M_1$ vs. $M_0$ and $M_2$ vs. $M_0$: $p=5.0 \times 10^{-2}$), despite the very low amount of data (51 polymorphic sites). Population genetics models appeared appropriate for the "Dsi_Begun" data set since they were not rejected by tests for goodness-of-fit.

DISCUSSION

The analysis of the "Dme_exhaustive" and "Dsi_exhaustive" data sets revealed a significant AT vs. GC asymmetry of non-coding polymorphism segregation in *D. melanogaster* and *D. simulans*, reflecting a biased transmission process and/or non-stationary base composition: GC→AT mutations are more frequent, or less likely to increase in frequency, or both, than AT→GC mutations, leading to an excess of low-AT-frequency polymorphisms. This effect is strong in GC-rich loci, but undetected in GC-poor loci. Normalized data sets were built to try and discriminate between the two (non-mutually exclusive) hypotheses. The biased transmission hypothesis explained allele frequency distribution in the "Dme_Glinka" data set significantly better than neutral models, either stationary or non-stationary. The fit of population genetics models to this data set was not good, however. This poor fit suggests that some of the assumptions underlying these models (e.g. panmixy, constant population size) are not met by the data, perhaps questioning the relevance of the likelihood-ratio tests. The "Dsi_Begun" data set revealed the same trends, but was smaller and little informative. These analyses, therefore, do not allow unambiguous

12
distinction between the "biased transmission" and "non-stationarity" hypothesis, although the former is best supported. We now discuss these two hypotheses in the light of our and other results.

Given that both D. melanogaster and D. simulans show an asymmetric segregation of AT vs. GC polymorphism, it is tempting to propose that this pattern resulted from a change of mutation bias prior to the split between these two lineages, in a neutral (i.e., unbiased transmission) context. This hypothesis, however, is not easy to reconcile with the correlation observed between AT vs. GC polymorphism asymmetry and GC-content – only GC-rich loci show a departure from the null hypothesis. It is difficult to imagine why a change in the mutational pattern should affect some, but not all regions of the genome. In addition, the mutation bias hypothesis would predict an accumulation of GC→AT substitutions. Kern and Begun (2004), however, recently reported that non-coding fixations were consistent with equilibrium base composition evolution in both D. melanogaster and D. simulans.

The hypothesis of a biased transmission of GC vs. AT alleles appears to accommodate the observed features more easily. The biased-transmission hypothesis predicts the strong correlation between asymmetric polymorphism and GC-content – loci undergoing a GC-biased segregation of polymorphisms tend to be GC-richer. Consistently, Singh et al. (2004) reported a significant correlation between recombination rate and GC substitution bias of a genome-wide repeated element in D. melanogaster. The GC-biased transmission model is also consistent with the stationary evolution of GC-content of non-coding sequences in D. melanogaster and D. simulans reported by Kern and Begun (2004).

We are concerned by the fact that the M2 hypothesis involves a constant population size in time. Analysis of sequence variation at synonymous sites of coding sequences, however, has revealed a reduction of selection efficacy for codon usage in the melanogaster lineage (Akashi 1996), leading to an accumulation of preferred →unpreferred substitutions,
and interpreted as the consequence of a recent drop of effective population size in *D. melanogaster* (Akashi 1996). This probable bottleneck should have affected the efficiency of GC-biased transmission in non-coding DNA as well, so that one would expect to observe an accumulation of GC→AT substitutions in *D. melanogaster* non-coding sequences. Why Kern and Begun (2004) did not report a significant asymmetry in the divergence pattern can be explained by a combination of several factors. First, the drop of population size might be too recent for a detectable accumulation of biased substitutions in non-coding sequences. Secondly, the strength of the transmission bias might be too weak, and the departure from equilibrium too small. Thirdly, the segregation bias reported here was detected from the GC-richest fraction of non-coding loci only, while Kern and Begun (2004) pooled all loci in their analysis.

It should be recalled that base composition shifts have been quite common in the Drosophila genus (Rodriguez-Trelles *et al.* 2000, Takano-Shimizu 2001, Tarrio *et al.* 2001). The mutation bias hypothesis, therefore, would require several changes of the mutation process in a relatively short period of time. Under the transmission bias model, however, such compositional shifts are expected when the effective population size changes, an event arguably more frequent than changes in the mutation process. For these various reasons, the most plausible model explaining recent GC-content evolution in non-coding sequences of *D. simulans* and *D. melanogaster* appears to be one in which (i) a GC-biased transmission process generally applies to some regions of the genome, leading to local increases in GC-content, and (ii) this process is interrupted, or its efficacy substantially reduced, when the effective population size declines, as it is suspected in *D. melanogaster*. Note that the base composition of non-coding DNA might also be affected by events of insertion and deletions. To account for the observed pattern, one may then assume that incoming insertions should be
GC-rich in general and deletions preferentially GC-poor. However, we have up to now no evidence for a bias in the base composition of insertions versus deletions.

One may be surprised that we could have detected a transmission bias from non-coding allele frequency patterns while comparable analyses applied to synonymous polymorphisms in *D. melanogaster* yielded equivocal results (Akashi 1997, Kliman 1999). If BGC was effective, one should expect a stronger signal from synonymous than from non-coding sites since synonymous sites undergo both BGC and selection for the GC-ending preferred codons. The number of sites examined in the above studies of synonymous polymorphism was much lower, however, than the 3508 + 648 used in the current analysis, reducing the power to detect biased segregation. Using a comprehensive *D. yakuba* / *D. simulans* / *D. melanogaster* data set, Kern and Begun (2004) obtained results essentially in agreement with ours. These authors, however, reported an intriguing difference between intron and intergenic sequence variation: introns, not intergenic sequences, showed a GC-biased segregation of polymorphisms in *D. melanogaster*. When we analysed introns and intergenic loci from the "Dme_Glinka" data set separately, we did not detect any difference between the two subsets of loci – both supported a GC-biased segregation.

Another difference between these studies and ours is that the former used to orient polymorphisms using *D. simulans* as an outgroup. We refrained from doing so in the current study. Orienting sites potentially increases the power to detect a segregation bias, but could strongly bias the analysis in case of a non-negligible amount of misorientation. A misoriented site will be seen, say, as a high-frequency AT→GC mutation when it is actually a low-frequency GC→AT one. To check this, we used the *D. simulans* outgroup available for the Dme_Glinka data set to orient A↔T and G↔C polymorphisms, thus avoiding biases related to the AT vs. GC mutation and fixation processes. We found a bimodal distribution of allele frequency, in which mutations with a frequency higher than 90% were 1.6 as numerous as
mutations with a frequency between 75% and 90%. One would expect a ratio of 0.58 under neutral evolution at demographic equilibrium, and an even lower proportion of high frequency mutations in the case of a bottleneck. We estimated that the proportion of misoriented mutations was higher than 5%, and probably around 10%, which makes a big difference when fitting population genetics models (e.g. see Baudry and Depaulis 2003), as we did, although this should not greatly affect the results of previous analyses (e.g. Kern and Begun 2004).

The transmission bias we detected, if confirmed, could be natural selection for G and C, or GC-biased gene conversion. Although our data does not allow discriminating between these two models, we tend to favour the BGC hypothesis over the selective one. This is because there is a large body of evidence demonstrating the effectiveness of BGC in yeast and mammals (Birdsell 2002, Galtier 2003, Kudla et al. 2004), whereas not a single example of selection on base composition in non-coding regions has been reported until now. Even hyperthermophilic prokaryotes, for which a putative advantage of C:G pairs at the DNA or RNA level could make sense, and in which population sizes are much higher than in Drosophila, show no evidence of selection on genomic GC-content (Galtier & Lobry 1997). The BGC model also accounts for the correlation between recombination rate and GC-content. BGC was recently invoked by Bartolome et al. (2004) to explain patterns of coding sequence variation in D. miranda. Whatever its origin, this transmission bias might explain the correlation between intron and exon GC-content. It also implies that the strength of selection for codon usage was previously overestimated, since a part of the AT↔GC asymmetry observed at synonymous sites must be a consequence of this bias.

Acknowledgments.
This work was supported by the CNRS Equipe Projet Multi Laboratoires "Méthodes informatiques pour la phylogénie moléculaire". We thank Adam Eyre-Walker and two anonymous referees for thoughtful comments on the manuscript.

Literature Cited.


Appendix: derivation of $f^*(i)$

The expected proportion of polymorphic sites consisting of $i$ C or G and $n-i$ A or T ($0<i<n$) in a sample of size $n$ under model M3 is given by:

$$f^*(i) = \pi(\alpha,W) \frac{Q_i(W)}{1-Q_0(W) - Q_n(W)} + (1-\pi(\alpha,W)) \frac{Q_{n-i}(-W)}{1-Q_0(-W) - Q_n(-W)}$$  \hspace{1cm} (A1)$$

The first part of equation (A1) considers polymorphic sites having arisen through an AT→GC mutations: $\pi(\alpha,W)$ is the expected proportion of such sites given distortion bias $W$ and mutation bias $\alpha$, and $Q_i(W)$ is the probability that a mutant allele with distortion bias $W$ shows frequency $i$ in the sample. Symmetrically, the second part of equation (A1) is for polymorphic sites having arisen through a GC→AT mutation. $\pi(W)$ in equation (A1) was calculated according to equation (3) in Smith and Eyre-Walker (2001), and $Q_i(W)$ is given by:

$$Q_i(W) = \int_0^1 D(x,W) P_i(x,W) dx$$  \hspace{1cm} (A2)$$

Here, $D(x, W)$ is the probability density of allele frequency $x$ in the population given $W$, and $P_i(x, W)$ is the probability that a site be at frequency $i$ in the sample given frequency $x$ in the population. $D(x, W)$ is obtained from standard population genetics (e.g. Sawyer and Hartl 1992):

$$D(x,W) = \frac{1 - e^{-W(1-x)}}{(1-e^{-W})x(1-x)}$$  \hspace{1cm} (A3)$$

$P_i(x, W)$ simply corresponds to a binomial sampling:

$$P_i(x,W) = \frac{n!}{i!(n-i)!} x^i (1-x)^{n-i}$$  \hspace{1cm} (A4)$$

The integration was performed numerically.
Table 1:

<table>
<thead>
<tr>
<th></th>
<th>lowGC</th>
<th>mediumGC</th>
<th>highGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>sites</td>
<td>443</td>
<td>444</td>
<td>362</td>
</tr>
<tr>
<td>GC</td>
<td>33.5%</td>
<td>39.4%</td>
<td>47.8%</td>
</tr>
<tr>
<td>$M_0$ lnL</td>
<td>-47.67</td>
<td>-59.69</td>
<td>-63.03</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.335</td>
<td>0.394</td>
<td>0.478</td>
</tr>
<tr>
<td>eqGC</td>
<td>33.5%</td>
<td>39.4%</td>
<td>47.8%</td>
</tr>
<tr>
<td>$M_1$ lnL</td>
<td>-45.67</td>
<td>-50.03</td>
<td>-44.30</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha^*$</td>
<td>0.275</td>
<td>0.259</td>
<td>0.243</td>
</tr>
<tr>
<td>eqGC</td>
<td>27.5%</td>
<td>25.9%</td>
<td>24.3%</td>
</tr>
<tr>
<td>$M_2$ lnL</td>
<td>-45.41</td>
<td>-48.40</td>
<td>-40.99</td>
</tr>
<tr>
<td>S*</td>
<td>0.3 [0.1-0.6]</td>
<td>0.7 [0.5-1]</td>
<td>1.2 [0.9,1.5]</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.272</td>
<td>0.244</td>
<td>0.216</td>
</tr>
<tr>
<td>eqGC</td>
<td>33.5%</td>
<td>39.4%</td>
<td>47.8%</td>
</tr>
<tr>
<td>$M_3$ lnL</td>
<td>-44.00</td>
<td>-46.29</td>
<td>-39.56</td>
</tr>
<tr>
<td>S*</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
</tr>
<tr>
<td>$\alpha^*$</td>
<td>0.250</td>
<td>0.226</td>
<td>0.213</td>
</tr>
<tr>
<td>eqGC</td>
<td>59.9%</td>
<td>59.1%</td>
<td>59.7%</td>
</tr>
</tbody>
</table>
Figure legends

Figure 1. Frequency distribution of the G or C allele in 3508 (respectively 648) AT vs GC polymorphisms from an exhaustive D. melanogaster (respectively D. simulans) data set. The left-most bar of the D. melanogaster histogram is for polymorphic sites with GC frequency lower than 0.1, the next bar for sites with GC frequency between 0.1 and 0.2, etc...

Figure 2. Asymmetry of GC→AT and AT→GC polymorphisms among loci in the "Dme_Glinka" data set. Each dot is for one locus. X-axis: locus GC-content. Y-axis: \( \frac{n_{AT} - n_{GC}}{n_{XY}} \), where \( n_{AT} \) is the number of AT↔GC polymorphic sites showing a frequency of the A or T lower than 50%, \( n_{GC} \) the number of AT↔GC polymorphic sites showing a frequency of the A or T higher than 50%, and \( n_{XY} \) the total number of polymorphic sites for the considered locus.

Figure 3. Expected allele frequency distribution under three bi-allelic population genetics models. Grey boxes: polymorphic sites having arised through an AT→CG mutation. White boxes: polymorphic sites having arised through a CG→AT mutation. Models M1 (non-stationary) and M2 (non-neutral) largely differ from M0 in that the expected distributions are asymmetric. Expectations under M1 and M2 are similar but distinct. For instance, with the parameter values chosen here for illustration, the less probable allele frequency is 5 under M1, but 6 under M2.

Figure 4. Observed allele frequency distribution for three subsets of the "Dme_Glinka" data set.
G or C "allele" frequency

figure 1
AT/GC polymorphism asymmetry
expected proportion of polymorphic sites

neutral, stationary ($M_0$)

neutral, non-stationary ($M_1$, $\alpha=0.35$)

non-neutral, stationary ($M_2$, $W=2.$)

expected proportion of polymorphic sites

G or C allele frequency

figure 3
number of polymorphic sites

GC-poor loci

GC-medium loci

GC-rich loci

number of polymorphic sites

G or C allele frequency

figure 4