biology letters

Biol. Lett. (2008) **4**, 438–441
doi:10.1098/rsbl.2008.0174
Published online 27 May 2008

Population genetics

Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*

Penelope R. Haddrill* **and Brian Charlesworth**

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King's Buildings, Edinburgh EH9 3JT, UK
**Author for correspondence (p.haddrill@ed.ac.uk).*

The nature of the forces affecting base composition is a key question in genome evolution. There is uncertainty as to whether differences in the GC contents of non-coding sequences reflect differences in mutational bias, or in the intensity of selection or biased gene conversion. We have used a polymorphism dataset for non-coding sequences on the X chromosome of *Drosophila simulans* to examine this question. The proportion of GC→AT versus AT→GC polymorphic mutations in a locus is correlated with its GC content. This implies the action of forces that favour GC over AT base pairs, which are apparently strongest in GC-rich sequences.

Keywords: GC content; biased gene conversion; selection; *Drosophila*

1. INTRODUCTION

A basic feature of an organism's genome is its nucleotide base composition, usually measured by the fraction of base pairs that are GC versus AT. This is highly variable among different parts of eukaryotic genomes. In particular, it tends to be reduced in regions of the genome with low levels of recombination, such as those around the centromeres (Díaz-Castillo & Golic 2007). Two main hypotheses have been proposed to explain such variation in GC content. The first involves differences in patterns of mutational bias. For selectively neutral sequences, the expected fraction of GC versus AT in a given region of the genome is determined by the ratio of the mutation rate for GC→AT to that for AT→GC; this ratio is the mutational bias parameter κ, (Sueoka 1962; Li 1987; Bulmer 1991). Differences in GC content can be caused by differences in κ, which can be estimated from patterns of nucleotide substitutions (Singh *et al.* 2005), and is generally larger than 1.

Alternatively, GC may be favoured over AT, owing to either natural selection, as with synonymous coding sequence sites (Akashi 1995), or biased gene conversion (BGC). BGC occurs when heterozygotes for GC and AT variants at a nucleotide site produce more than 50% of the GC variant in their gametes, as a result of biased repair of DNA heteroduplexes (Marais 2003). BGC causes an expected change in the frequency of GC versus AT variants at a site similar to that caused by selection (Gutz & Leslie 1976). The greater the intensity of selection or BGC in favour of GC, compared with mutation and genetic drift, the higher the equilibrium GC content of a sequence (Li 1987; Bulmer 1991).

Data on both interspecies divergence and within-species polymorphism permit the detection of selection/BGC, since these forces are less effective at preventing disfavoured variants (AT in this case) entering the population as polymorphic variants than at preventing them becoming fixed (Akashi 1995). If these forces are acting, we should therefore see more GC→AT relative to AT→GC variants among polymorphisms, compared with substitutions between species. Equilibrium for base composition implies an equal number of GC→AT and AT→GC substitutions along a lineage, regardless of the action of selection or BGC. An excess of GC→AT over AT→GC for polymorphisms then indicates the action of selection/BGC (Akashi 1995).

This allows the estimation of the intensity of selection or BGC per site (multiplied by four times the effective population size, N_e) from the proportion of GC→AT polymorphisms among GC→AT and AT→GC polymorphisms (Maside *et al.* 2004). This scaled estimate of selection/BGC is denoted by γ. Other methods for estimating γ use information on the frequency distribution of variants in the population (Akashi 1999; Galtier *et al.* 2006). A difficulty is that the assumption of equilibrium is often violated; this is known to be the case, for example, for both *Drosophila melanogaster* (Akashi *et al.* 2006) and humans (Duret *et al.* 2006).

Here, we present an analysis of a dataset on polymorphisms in non-coding sequences in a sample of *Drosophila simulans* from Madagascar, together with estimates of divergence from their homologues in *Drosophila melanogaster* and *Drosophila yakuba*. We detect the signature of selection/BGC, especially for sequences with high GC content.

2. MATERIAL AND METHODS

(a) *Source of data*

We used a total of 44 X-linked non-coding loci from the dataset of Haddrill *et al.* (in press), including 23 introns, ten 5′ untranslated transcribed regions (UTRs) and eleven 3′UTRs. Each locus was surveyed in a sample of 20 *D. simulans* males from the putatively ancestral Madagascan population (Dean & Ballard 2004) and was aligned with the homologous *D. melanogaster* (http://flybase.org/, release 4.2) and *D. yakuba* (http://insects.eugenes.org/species/blast) sequences, as described in Haddrill *et al.* (in press).

(b) *Data analysis*

To polarize the origin of polymorphisms within *D. simulans*, and to determine the fixed differences between *D. simulans* and *D. melanogaster* that occurred along the *D. simulans* lineage, we reconstructed a *D. melanogaster–D. simulans* ancestral sequence using *D. yakuba* as an out-group and estimated the number of polymorphisms and substitutions, as described in Haddrill *et al.* (in press). This allows us to classify both polymorphisms and substitutions along the *D. simulans* lineage as GC→GC, AT→AT, GC→AT or AT→GC. Only the latter two classes are of interest for our purposes. The results for each locus are presented in the electronic supplementary material 1. We estimated γ using the maximum-likelihood method of Maside *et al.* (2004).

3. RESULTS

We tested for base composition equilibrium by examining the pattern of GC→AT versus AT→GC

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2008.0174 or via http://journals.royalsociety.org.

Received 27 March 2008
Accepted 8 May 2008

Table 1. Numbers of GC→AT and AT→GC substitutions and polymorphisms in different classes of sequences. (High, medium and low GC content sequences correspond to loci with GC contents in the ranges 43–56% ($n=15$, mean 45%), 37–42% ($n=15$, mean 39%) and 26–36% ($n=14$, mean 34%), respectively.)

class of sequence	substitutions GC→AT	substitutions AT→GC	polymorphisms GC→AT	polymorphisms AT→GC
intron	37	44	265	231
5′UTR	24	24	81	55
3′UTR	11	23	47	42
GC content				
high	19	22	119	76
medium	32	38	136	119
low	21	31	138	133

substitutions along the branch of the phylogeny leading to *D. simulans* from its common ancestor with *D. melanogaster* (see §2). Table 1 shows the numbers of different types of substitutions inferred for the three classes of non-coding DNA. Introns and 5′UTR sequences show no significant departure from the 1 : 1 ratio of GC→AT versus AT→GC substitutions expected under equilibrium base composition; there is a marginally significant ($\chi^2=4.24$, $p<0.05$) deficit of GC→AT substitutions for 3′UTR sequences. For the pooled dataset, there is no significant departure from equality ($\chi^2=2.21$), and there is no significant heterogeneity among the three classes of sequence. If we divide the set of loci into three nearly equal-sized classes with respect to their GC contents, then there are no significant differences among the high, medium and low GC content loci. Overall, there is no evidence for an excess of GC→AT over AT→GC substitutions, consistent with Akashi *et al.* (2006).

We then asked whether there is an excess of GC→AT over AT→GC mutations among polymorphisms. If base composition is close to equilibrium, this is an indicator of selection or BGC in favour of GC base pairs (see §1). Table 1 shows the numbers of polymorphic variants in the three different classes of non-coding sequence. For introns and 3′UTR sequences there is no significant departure from 1 : 1 ($\chi^2=2.33$ and 0.29, respectively), whereas for 5′UTR sequences, $\chi^2=4.97$ ($p<0.02$; all p values for tests of 1 : 1 ratios are one tailed). If all sequences are pooled, $\chi^2=5.86$ ($p<0.01$); this result is not simply due to the 5′UTR sequences alone, since we find no significant heterogeneity between the 5′UTR sequences and the introns and 3′UTR sequences combined ($\chi^2=1.72$, $p>0.05$). This suggests that GC is favoured over AT.

The three classes of non-coding sequences differ in their GC content, with means of 37, 40 and 50% for intronic, 3′UTR and 5′UTR sequences, respectively. If we pool the three classes, and divide the data into high, medium and low GC content sequences as above, the χ^2 values for 1 : 1 for polymorphisms in each category are 9.48, 1.33 and 0.09, respectively. The first of these has $p<0.001$. The heterogeneity χ^2 between these categories is 4.84 ($p<0.05$). The proportion of GC→AT among GC→AT and AT→GC polymorphisms at a locus is significantly correlated with its GC content (figure 1*a*).

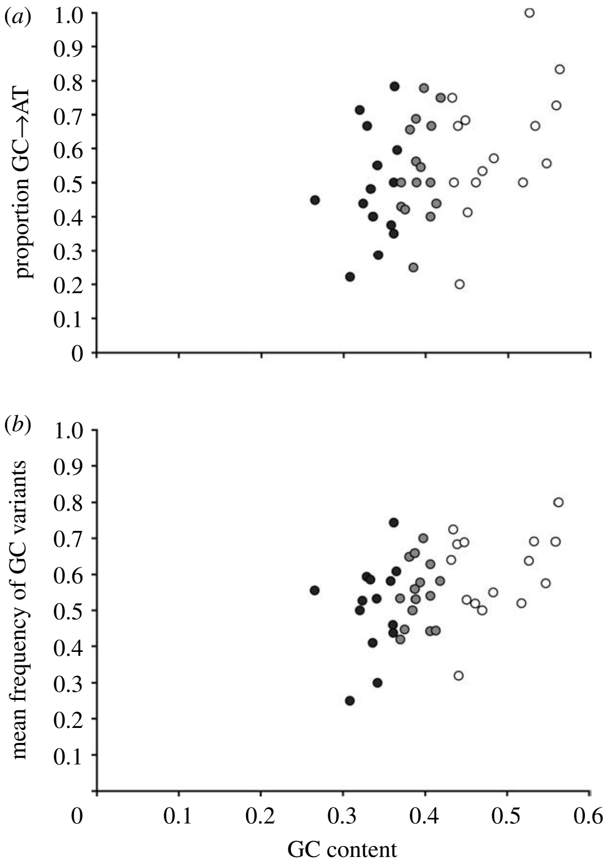

Figure 1. Relationships between the GC content of a locus and (*a*) the proportion of GC→AT among GC→AT and AT→GC polymorphisms and (*b*) the mean frequency of GC variants. The Spearman's rank correlations are 0.377 ($p<0.01$) and 0.355 ($p<0.01$), respectively. Black circles, low GC; grey circles, medium GC; white circles, high GC.

This suggests that there is a tendency for GC-rich sequences to be associated with stronger selection/BGC in favour of GC. We investigated this by estimating the scaled selection parameter γ (see §2). We first fitted a common γ to all the loci, obtaining a maximum-likelihood estimate of $\gamma=0.25$, $\ln L=-496.83$, with two-unit support limits 0.04–0.47. Table 2 shows the results of fitting separate γ values to the three categories of GC content, with strong support for a positive γ only for the high GC content sequences (the lower three-unit support limit in this case is 0.13). The difference in fit between the models with a common γ and with individual values fitted is significant ($\chi_2^2=5.04$, $p<0.02$ on a one-tailed test).

Table 2. Maximum-likelihood estimates of the scaled selection parameter γ

GC content category	maximum likelihood γ	maximum log-likelihood	two-unit support limits
high	0.60	−130.38	(0.21–1.00)
medium	0.20	−176.19	(−0.16–0.54)
low	0.05	−187.80	(−0.35–0.38)

If GC-rich sequences are associated with stronger selection/BGC in favour of GC, we would expect to see GC variants present at a higher mean frequency in this category, compared with the medium and low GC content categories (Galtier *et al.* 2006). For each category, we summed the number of variants in the GC and AT states across every GC/AT polymorphic site, and compared the results with the neutral expectation of equal numbers. All three categories show an excess of variants in the GC state compared with the AT state (GC contents: high, $\chi^2 = 173.25$; medium, $\chi^2 = 33.93$; low, $\chi^2 = 31.93$; all $p < 0.001$), and a test of heterogeneity indicates that this deviation differs between categories ($\chi^2 = 49.36$, $p < 0.001$). We also find a positive correlation between the mean frequency of GC variants for a locus and its GC content (figure 1*b*), as expected if GC-rich sequences are associated with stronger selection/BCG in favour of GC. Given the potential biases associated with ancestral inference (Eyre-Walker 1998; Akashi *et al.* 2007), we can also use these unpolarized data to recalculate γ, using the method of Cutter & Charlesworth (2006). Estimates are not significantly different from those reported above (see the electronic supplementary material 2), indicating that inference biases do not account for the patterns that we see.

4. DISCUSSION

Our results suggest that non-coding sequences with high GC contents are associated with stronger selection/BGC in favour of GC than sequences with low or intermediate GC contents. It is of interest to compare the observed GC contents with the equilibrium values predicted from the Li–Bulmer equation, $1/(1+\kappa \exp -\gamma)$ (Li 1987; Bulmer 1991). The results of Singh *et al.* (2005) suggest an estimate of κ of 2.1 for non-functional sequences in heterochromatic regions with very low GC content in *D. melanogaster*; these are the least likely to be affected by gene conversion, although it may not be completely absent (Gay *et al.* 2007). If we use this value of κ in conjunction with the maximum-likelihood estimates of γ in table 2, the predicted values of GC contents are 33, 37 and 46%, for the low, medium and high GC sequences, respectively. These agree with the mean GC contents for these regions (table 1). Use of the standard formula for fixation probabilities of mutations affected by selection (Kimura 1983, p. 43) shows that there is approximately only a 5% underestimation of κ from substitution patterns, if we use the maximum-likelihood estimate of γ for low GC content from table 2. A κ of 2.1 thus seems to fit the data well, and our estimates of γ are consistent with the hypothesis that differences in GC content among non-coding sequences reflect differences in the intensity of selection or BGC in favour of GC. Selection in favour of GC would imply functionality of GC base pairs in non-coding sequences, but there is currently no way to distinguish between selection and BGC using these data.

These results contrast with those of Galtier *et al.* (2006) who used a different method to analyse a *D. melanogaster* non-coding polymorphism dataset for an African population (Glinka *et al.* 2003). Their best-fitting model gave γ and κ estimates of between 1.5 and 1.7 and between 3 and 3.7, respectively, for low, medium and high GC content sequences. These predict a GC content of 59%, very different from the observed values. The most likely explanation of these discrepancies is that this population is not at statistical equilibrium, for which there is other evidence (Li & Stephan 2005). Indeed, it is unlikely that γ for non-coding sequences could be as high as the estimates of Galtier *et al.* (2006), since values for synonymous sites are of the order of 2 in several *Drosophila* species (Maside *et al.* 2004; Bartolomé *et al.* 2005; Comeron & Guthrie 2005), and these include the effects of both BGC and selection on codon usage bias. The same reservation applies to the estimates of γ obtained for humans by Lercher *et al.* (2002).

We are very grateful to Peter Andolfatto and Doris Bachtrog for allowing us access to their library of perl scripts (‘Polymorphorama’), which we used to analyse this data, and two anonymous reviewers for their comments on the manuscript. P.R.H. was supported by grants from the BBSRC and NERC to B.C., and B.C. was supported by the Royal Society.

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

Akashi, H. 1999 Inferring the fitness effects of DNA polymorphisms and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**, 221–238.

Akashi, H., Ko, W.-Y., Piao, S., John, A., Goel, P., Lin, C.-F. & Vitins, A. 2006 Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172**, 1711–1726. (doi:10.1534/genetics.105.049676)

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* **10**, e1065. (doi:10.1371/journal.pone.0001065)

Bartolomé, C., Maside, X., Yi, S., Grant, A. L. & Charlesworth, B. 2005 Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*. *Genetics* **169**, 1495–1507. (doi:10.1534/genetics.104.033068)

Bulmer, M. G. 1991 The selection–mutation–drift theory of synonomous codon usage. *Genetics* **129**, 897–907.

Comeron, J. M. & Guthrie, T. B. 2005 Intragenic Hill–Robertson interference influences selection on synonymous mutations in *Drosophila*. *Mol. Biol. Evol.* **22**, 2519–2530. (doi:10.1093/molbev/msi246)

Cutter, A. D. & Charlesworth, B. 2006 Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Curr. Biol.* **16**, 1–5. (doi:10.1016/j.cub.2006.08.067)

Dean, M. D. & Ballard, J. W. 2004 Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Mol. Phylogenet. Evol.* **32**, 998–1009. (doi:10.1016/j.ympev.2004.03.013)

Díaz-Castillo, C. & Golic, K. G. 2007 Evolution of gene sequence in response to chromosomal location. *Genetics* **177**, 359–374. (doi:10.1534/genetics.107.077081)

Duret, L., Eyre-Walker, A. & Galtier, N. 2006 A new perspective on isochore evolution. *Gene* **385**, 71–74. (doi:10.1016/j.gene.2006.04.030)

Eyre-Walker, A. 1998 Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**, 686–690. (doi:10.1007/PL00006427)

Galtier, N., Bazin, E. & Bierne, N. 2006 GC-biased segregation of non-coding polymorphisms in *Drosophila*. *Genetics* **172**, 221–228. (doi:10.1534/genetics.105.046524)

Gay, J., Myers, S. & McVean, G. 2007 Estimating meiotic gene conversion rates from population genetic data. *Genetics* **177**, 881–894. (doi:10.1534/genetics.107.078907)

Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. 2003 Demography and selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**, 1269–1278.

Gutz, H. & Leslie, J. F. 1976 Gene conversion: a hitherto overlooked parameter in population genetics. *Genetics* **83**, 861–866.

Haddrill, P. R., Bachtrog, D. & Andolfatto, P. In press. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol. Biol. Evol.* (doi:10.1093/molbev/msn125)

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

Lercher, M. J., Smith, N. G. C., Eyre-Walker, A. & Hurst, L. D. 2002 The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**, 1805–1810.

Li, W.-H. 1987 Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345. (doi:10.1007/BF02134132)

Li, H. P. & Stephan, W. 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* **171**, 377–384. (doi:10.1534/genetics.105.041368)

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

Maside, X., Weishan Lee, A. & Charlesworth, B. 2004 Selection on codon usage in *Drosophila americana*. *Curr. Biol.* **14**, 150–154. (doi:10.1016/j.cub.2003.12.055)

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

Sueoka, N. 1962 On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592. (doi:10.1073/pnas.48.4.582)