Genet. Res., Camb. (2003), **81**, *pp.* 79–87. *With 2 figures.*
DOI: 10.1017/S0016672302006079 Printed in the United Kingdom

Neutral effect of recombination on base composition in *Drosophila*

GABRIEL MARAIS*, DOMINIQUE MOUCHIROUD AND LAURENT DURET
Laboratoire 'Biométrie et Biologie Evolutive', UMR 5558, Bâtiment Gregor Mendel, Université Claude Bernard – Lyon 1, 69622 Villeurbanne Cedex, France

(*Received 23 September 2002 and in revised form 13 November 2002*)

Summary

Recombination is thought to have various evolutionary effects on genome evolution. In this study, we investigated the relationship between the base composition and recombination rate in the *Drosophila melanogaster* genome. Because of a current debate about the accuracy of the estimates of recombination rate in *Drosophila*, we used eight different measures of recombination rate from recent work. We confirmed that the G+C content of large introns and flanking regions is positively correlated with recombination rate, suggesting that recombination has a neutral effect on base composition in *Drosophila*. We also confirmed that this neutral effect of recombination is the main determinant of the correlation between synonymous codon usage bias and recombination rate in *Drosophila*.

1. Introduction

In many species, including *Drosophila melanogaster*, codon bias is mainly shaped by weak selection to enhance translation efficiency (Shields *et al*., 1988; Hartl *et al*., 1994; Akashi, 1995; Moriyama & Powell, 1997; Duret & Mouchiroud, 1999). Nevertheless, in such species, it is also recognized that neutral processes affecting base composition (e.g. mutation pressure) can partly be responsible for variation in codon bias across the genome (Kliman & Hey, 1994; Akashi *et al*., 1998). In *D. melanogaster*, codon bias has been found to increase weakly but significantly with recombination rate (Kliman & Hey, 1993; Comeron *et al*., 1999; Marais *et al*., 2001; Hey & Kliman, 2002). Three models (one selective and two neutral) are currently proposed to explain this observation. The selective model proposes that the positive correlation between codon bias and recombination rate is due to Hill–Robertson interference (HRi). HRi is a well-known population genetics concept that corresponds to a decrease of selection efficacy because of genetic linkage in finite populations (Hill & Robertson, 1966; Felsenstein, 1974). The neutral models stem from the observation that, in *D. melanogaster*, most (21/22) of the optimal codons end in G or C (Shields *et al*., 1988; Duret & Mouchiroud, 1999). Thus, the high frequency of optimal codons observed in regions of high recombination might be a consequence of a GC-biased mutation pressure (Perry & Ashworth, 1999) or of GC-biased gene-conversion events in those regions (Galtier *et al*., 2001; Birdsell, 2002). According to the selective model, only synonymous sites undergoing selection on codon usage should be affected by recombination rate. By contrast, the neutral models predict that all silent sites (both synonymous sites and unconstrained non-coding DNA) should be affected by the recombination rate.

Thus, a simple test of these models consists in investigating the correlation between the base composition of non-coding DNA (introns, flanking regions) and the recombination rate. It is likely that such regions are not totally neutral, notably because of the presence of regulatory elements (promoters, enhancers, splice signals etc.). However, they are generally relatively weakly constrained, and hence – although they do not exactly reflect the neutral substitution pattern – they can reveal variation in the neutral substitution pattern. We found that the G+C content of large introns and intergenic regions correlates positively with the recombination rate in the complete

* Corresponding author. Tel: +44 (0) 131 650 5543. Fax: +44 (0) 131 650 6564. e-mail: Gabriel.Marais@ed.ac.uk
Present address: Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK.

Table 1. *Description of the different estimates of recombination rate in* D. melanogaster

	Estimation method	Genetic markers	Physical map	References
ACE	Coefficient of exchange	–	Cytogenetic	Kindhal, 1994; Hey & Kliman, 2002
CC99	Sliding window; WS = 9 bands	–	Cytogenetic	Carvalho & Clark, 1999
CK00	Polynomial; DP not mentioned	–	Cytogenetic	Comeron & Kreitman, 2000; Comeron *et al*., 1999
HK-p	Polynomial; DP = 4 for each arm	493	CG2	Hey & Kliman, 2002
HK-w	Sliding window; WS = 8 markers	493	CG2	Hey & Kliman, 2002
KH93	Polynomial; DP = 4–5 per chromosome	–	Cytogenetic	Kliman & Hey, 1993; Hey & Kliman, 2002
MMD01	Polynomial; DP = 2 for each arm	892	CG1	Marais *et al*., 2001
RTE	Sliding window; WS = 8 markers	–	Cytogenetic	Hey & Kliman, 2002

CG1, first release of the complete genome (24th March 2000); CG2, second release of the complete genome (18th October 2000); DP, degree of polynomial curves; WS, window size; –, number of genetic markers not mentioned.

genome of *D. melanogaster* (Marais *et al*., 2001). Therefore, our data agree with the neutral models and not with the selective one. Recently, a study conducted on a similar dataset led to the conclusion that there is no correlation between the G+C content of non-coding DNA and the recombination rate in the complete genome of *D. melanogaster* (Hey & Kliman, 2002). Therefore, the authors rejected the neutral models and retained the selective one. They explained the discrepancy between our conclusions by a methodological artefact in our study. We estimated recombination rate by fitting second-degree polynomial curves for each chromosome arm (Marais *et al*., 2001). Hey & Kliman (2002) proposed that these estimates are erroneous and generate a positive correlation between the recombination rate and the G+C content of non-coding DNA by assigning high recombination rates to telomeric regions, which do indeed have high G+C content but actually have reduced levels of recombination.

Another possible explanation for the discrepancy between the results of the two studies comes from the way in which the G+C content of the non-coding DNA is measured. Hey & Kliman (2002) determined the G+C content of non-coding DNA (*GCnc*) from intron G+C content and, for those genes without introns (22% of the data set), from the immediately flanking non-coding DNA (at most 1000 bp on either side). However, the distribution of G+C content is significantly different between introns and flanking regions (introns have an average G+C content of 37%, 5′ flanking regions have an average G+C content of 39%, and 3′ flanking regions have an average G+C content of 36%, Wilcoxon's tests with $p < 10^{-4}$; see also Fig. 2 in Marais *et al*., 2001). Hence, if there is a weak dependence of G+C content on recombination rate in introns and flanking regions, *GCnc* might not capture it. Moreover, *GCnc* was calculated from all introns, including very short ones. In *D. melanogaster*, 60% of introns are less than 100 bp long (Adams *et al*., 2000). Within such short introns, a significant proportion of the sequence is strongly constrained because of the presence of elements required for the splicing reaction (the total length of splice donor, acceptor and branch-point consensus signals is 27 bp) (Mount *et al*., 1992). Because *GCnc* includes a large amount of constrained non-coding DNA and mixes introns and flanking regions, the test of the selective and neutral models by Hey & Kliman (2002) might be inaccurate. We measured the correlations between recombination rate and non-coding DNA G+C content independently for introns, 5′ and 3′ flanking regions (Marais *et al*., 2001). Moreover, for each correlation, we only retained genes for which the length of non-coding sequences was at least 200 bp.

2. Objectives and methods

Here, we investigated whether our different conclusions are due to a methodological artefact in the measures of recombination rate in Marais *et al*. (2001) or to the way the G+C content of the non-coding DNA is measured in Hey & Kliman (2002). We considered eight different estimates of recombination rates that have been used in recent work (Carvalho & Clark, 1999; Comeron *et al*., 1999; Marais *et al*., 2001; Hey & Kliman, 2002) (Table 1). In all cases, recombination rate is estimated by using Marey maps: the genetic positions (in centiMorgans, cM) and physical positions (in megabases) of markers that

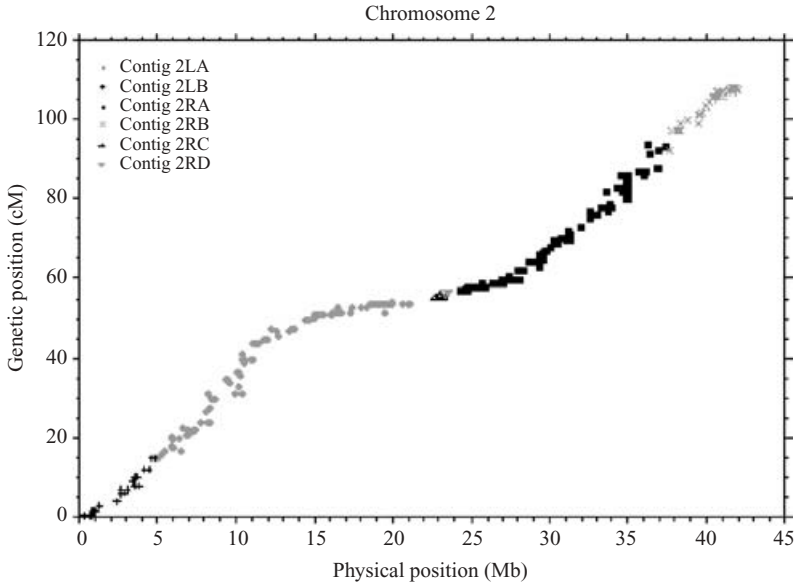

Fig. 1. Marey map of the second chromosome of *D. melanogaster*. In this chromosome, 317 markers have been mapped in both the genetic and the physical maps. Sequence data come from the second release of the complete genome of *D. melanogaster* (October 2000). The sequence of chromosome 2 was available in six large contigs (indicated by dots of different colours). Gaps with unknown size between those contigs are not included.

have been localized on both kinds of maps are plotted (Fig. 1). The recombination rate (cM/Mb) at a given position of the chromosome is given by the slope of the curve at that position. The approaches differ in the ways that the slope is measured. A first possibility is to fit a polynomial curve to the whole dataset for each chromosome arm. The derivative of this function is then used to estimate the local recombination rate. Another approach uses sliding windows along the chromosome: within each window, a linear function is fitted to the data, and the slope of this line is taken as the estimate of the local recombination rate. Figure 2 shows that there are large differences between these estimates, although they all co-vary, as Hey & Kliman (2002) also noticed. Various parameters affect the estimation of recombination rates, including the nature of the physical map, the degree of the polynomial curve and the size of the sliding window. It is not yet clear which of these approaches is the most accurate. Estimating recombination by the polynomial method has a smoothing effect and probably obscures regional variation that can be detected with sliding windows; for example, the polynomial method tends to overestimate recombination rates in telomeric regions. Conversely, the sliding-window approach is more sensitive to errors in the genetic or physical locations, and hence might generate artefactual variations in recombination rate. Estimates published before 2000 are probably less reliable than more recent ones, because they used physical maps based on cytogenetic data, which are less accurate than the nearly complete genome sequence that are now available.

3. Results

(i) *Recombination rate and non-coding DNA G+C content*

Table 2 shows the correlations between non-coding DNA G+C content and the different estimates of recombination rates. When all introns are used, only MMD01 correlates positively with intron G+C content, in agreement with the results of Hey & Kliman (2002). However, as mentioned above, the base composition of short introns is not neutral and these introns might not be an accurate indicator of neutral substitution patterns across the genome. When only introns larger than 100 bp are selected, six out of eight estimates of recombination rate correlate significantly with intron G+C content, and always positively. When telomeric regions (for which the estimates of recombination rate might be inaccurate) are excluded, seven out of eight estimates of recombination rate correlate positively with intron G+C content.

In both 5′ and 3′ flanking regions, seven out of eight estimates of recombination rate are significantly and positively correlated with the G+C content (Table 2). However, the correlations measured in 5′ flanking regions are generally weaker than in introns or 3′ flanking regions, and are more sensitive to the noise introduced by telomeric regions: the number of significant correlations (always positive) increase from three to seven out of eight when telomeric regions are excluded. In general, promoter elements are located upstream of genes. Thus, 5′ flanking regions might be more constrained than other non-coding regions, which would explain the weakness of the correlations

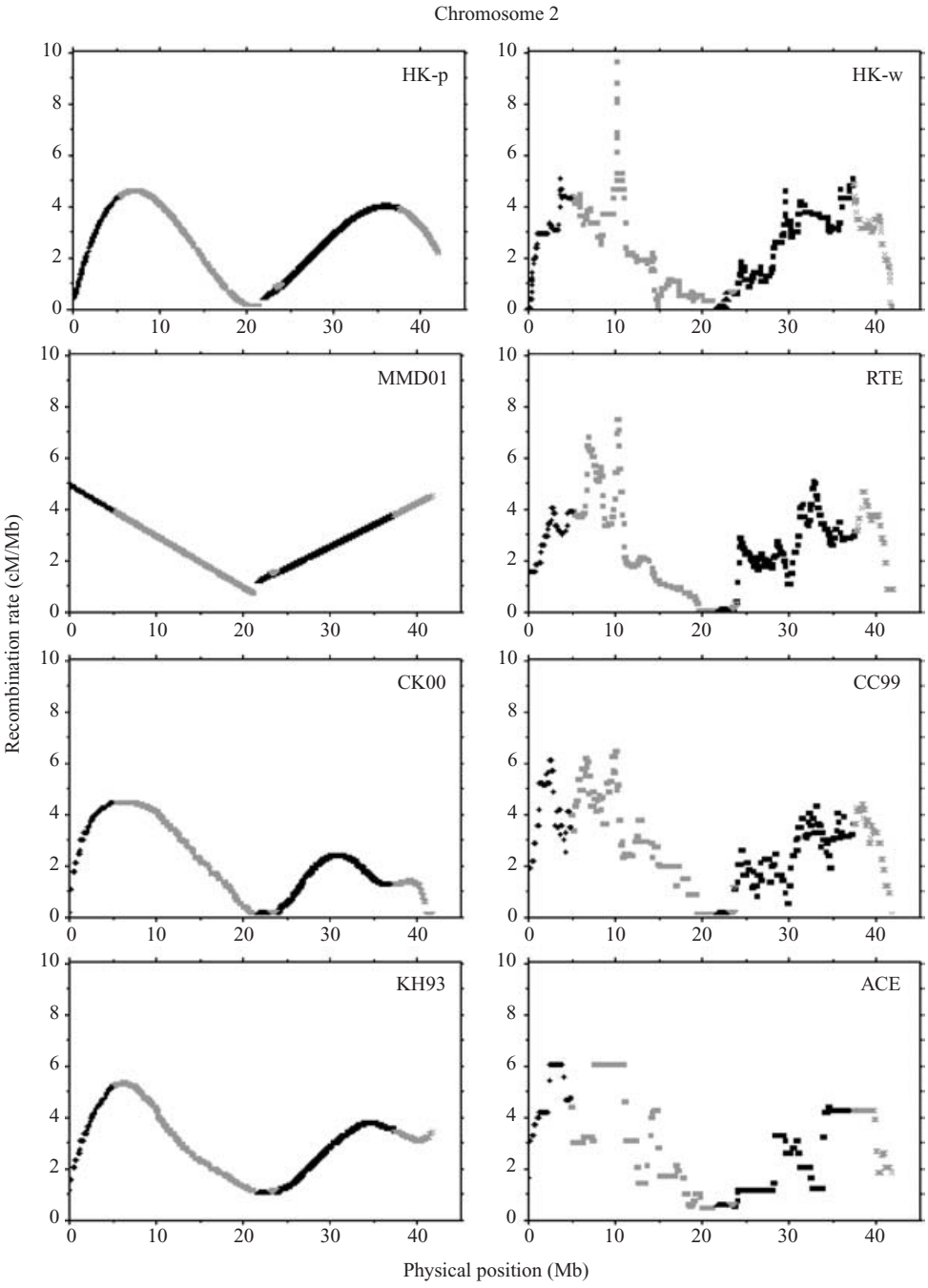

Fig. 2. Comparison of the different estimates of recombination rate in the second chromosome of *D. melanogaster*. The sequence of chromosome 2 was available in six large contigs (indicated by dots of different colours). Gaps with unknown size between those contigs are not included.

detected. Despite this specificity of the 5′ flanking regions, it should be noted that the overall results for flanking regions and introns are consistent with a positive correlation between the G+C content of non-coding DNA and the recombination rate.

Thus, the first prediction of the neutral models is confirmed here with different measures of recombination rate, including those of Hey & Kliman (2002). In conclusion, the discrepancy between the results of Marais *et al*. (2001) and Hey & Kliman (2002) is not due to the difference between measures of recombination rate but, instead, to the fact that Hey & Kliman (2002) included very short introns and mixed introns and flanking regions to calculate the G+C content of non-coding DNA.

(ii) *Recombination rate and the frequency of non-optimal codons*

Another prediction of the selective and neutral models concerns the frequency of the non-optimal codons. According to the neutral models, recombination

Table 2. *Correlation between the G + C content of non-coding DNA and the different estimates of recombination rate in* D. melanogaster. *The G + C content of introns and flanking regions were measured using sequence data from the 19 large contigs of the second release of the complete genome of* D. melanogaster (*Adams* et al.*, 2000*) (*http://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/Scaffolds/LARGE*). *KH93, ACE, RTE, HK-p and HK-w recombination data come from the spreadsheet provided by Hey & Kliman* (*2002*) (*http://lifesci.rutgers.edu/~heylab*). *A total of 12 266 genes matched in both datasets, which is close to the 12 999 genes used in Hey & Kliman* (*2002*). *Telomeric regions correspond to cytogenetic bands 1, 21, 60–61 and 100* (*Kliman & Hey, 1993*)

	N^a	MMD01[b]	KH93[b]	ACE[b]	RTE[b]	HK-p[b]	HK-w[b]	CC99[b]	CK00[b]
All introns	9563 (8266)	0·150***	0·010	0·005	0·015	−0·004	0·019	0·003	−0·040***
Introns >100 bp[c]	6263 (5322)	0·177***	0·047***	0·024	0·046***	0·035*	0·050***	0·040**	−0·020
Introns >100 bp[c] without telomeric regions	5918 (5040)	0·143***	0·060***	0·048***	0·077***	0·056***	0·071***	0·060***	0·013
5′ Flanking regions[d]	9436 (8340)	0·139***	0·030**	0·021*	0·017	0·008	0·016	0·014	−0·014
5′ Flanking regions without telomeric regions[d]	11 369 (10 023)	0·112***	0·038***	0·035**	0·039***	0·023*	0·029*	0·026*	0·006
3′ Flanking regions[d]	9151 (8087)	0·144***	0·037***	0·059***	0·040***	0·027*	0·043***	0·039***	−0·014
3′ Flanking regions without telomeric regions[d]	8625 (7625)	0·131***	0·044***	0·068***	0·057***	0·035**	0·051***	0·050***	0·011

[a] Values in parenthesis correspond to sample size for MMD01.
[b] Values correspond to the Spearman's coefficient of correlation (*Rs*). ***, $p < 0{\cdot}0005$; **, $p < 0{\cdot}005$; *, $p < 0{\cdot}05$; all other values are not significant.
[c] In genes containing more than one intron, all introns larger than 100 bp were concatenated.
[d] The 5′ and 3′ flanking regions of a gene are defined here as sequences, spanning at most 1000 bp upstream of the start codon and downstream of the stop codon, respectively, up to the extremity (start or stop codons) of the neighbouring gene. 5′ and 3′ flanking regions smaller than 100 bp were excluded. We also excluded overlapping 3′ and 5′ flanking regions (22% of the dataset). The remaining 78% correspond to flanking regions extracted from intergenic regions larger than 2000 bp or to 3′ flanking regions in configuration 3′–3′ (flanking genes in head-to-head orientation) and the 5′ flanking regions in configuration 5′–5′ (flanking genes in tail-to-tail orientation). Including the overlapping 3′ and 5′ flanking regions in the analysis gives similar results.

Table 3. *Correlation between the frequency of non-optimal codons and the different estimates of recombination rate in* D. melanogaster. *Non-optimal codons correspond to the codons whose frequency does not increase in highly expressed genes, namely AGG, CGG, AGC, ACG, CCG, GCG, GGG, UUG, AGA, CGA, CUA, CUU, UUA, AGU, UCA, ACA, ACU, CCA, CCU, GCA, GCU, GGA, GUU, AAA, AAU, CAA, CAU, GAA, GAU, UAU, UGU, UUU, AUA, AUU, UCU, GGU, GUA* (*Duret & Mouchiroud, 1999*). Fnop-GC *and* Fnop-AU *data come from the spreadsheet provided by Hey & Kliman* (2002) (*http://lifesci.rutgers.edu/~heylab*)

	N^a	MMD01[b]	KH93[b]	ACE[b]	RTE[b]	HK-p[b]	HK-w[b]	CC99[b]	CK00[b]
Fnop-GC	12 266 (10 766)	0·055***	−0·005	0·028**	0·014	0·003	0·014	0·011	0·000
Fnop-GC without telomeric regions	11 539 (10 144)	0·047***	0·000	0·033***	0·023*	0·008	0·020*	0·017	0·009
Fnop-AU	12 266 (10 766)	−0·138***	−0·040***	−0·044***	−0·061***	−0·038***	−0·062***	−0·034***	−0·019*
Fnop-AU without telomeric regions	11 539 (10 144)	−0·131***	−0·040***	−0·047***	−0·066***	−0·038***	−0·056***	−0·036***	−0·028**

[a] Values in parenthesis correspond to sample size for MMD01.
[b] Values correspond to the Spearman's coefficient of correlation (*Rs*). ***, $p<0·0005$; **, $p<0·005$; *, $p<0·05$; all other values are not significant.

should increase the G+C content of all silent DNA. Consequently, the frequency of AU-ending non-optimal codons (*Fnop-AU*) should decrease with recombination, whereas the frequency of GC-ending non-optimal codons (*Fnop-GC*) should increase. Conversely, according to the selective model, recombination should improve the efficacy of selection against all non-optimal codons. Consequently, the frequency of non-optimal codons should decrease with recombination whatever their ending base.

Table 3 shows the correlations between *Fnop* and different estimates of recombination rate. In our previous analysis (Marais *et al*., 2001), we found that the recombination rate was negatively correlated with *Fnop-AU* and positively correlated with *Fnop-GC*. The negative correlation with *Fnop-AU* is confirmed here with most of the measures of recombination rate. A positive correlation with *Fnop-GC* is detected with four measures of recombination rate (MMD01, ACE, RTE and HK-w) when telomeric regions are excluded but, for the remaining measures of recombination, no negative correlation with *Fnop-GC* has been detected. Hence, there is no evidence that recombination increases the efficacy of selection against non-optimal codons. On the contrary, for the four cases in which a significant correlation is found, this correlation is positive. This test thus supports the neutral models rather than the selective model.

4. Discussion

(i) *Estimates of recombination rate*

There are some significant differences between the estimates of recombination rate obtained by different methods (Fig. 2). As mentioned previously, there is not yet a consensus about which is the most reliable method. However, despite this methodological limit, most measures reveal similar relationships between recombination rates and codon bias or non-coding DNA G+C content (Tables 2, 3). Thus, the discrepancy between the recent studies on the topic is not due to differences between recombination rate estimates, notably in telomeric regions. Indeed, as mentioned previously (Marais *et al*., 2001) and confirmed here (Tables 2, 3), removing the telomeric regions does not affect the results.

Interestingly, MMD01 is clearly the method that produces the highest correlation coefficients with codon bias and non-coding DNA G+C content (Tables 2, 3). How can we explain this observation? The overestimation of recombination rates in telomeric regions with MMD01 does not explain it because, when these regions are removed, the correlation coefficients between MMD01 and codon bias or non-coding DNA G+C content remain the highest (Tables 2, 3). Compared with other methods,

MMD01 is the one that has the strongest smoothing effect (Fig. 2). Thus, a possible explanation is that MMD01, by underestimating regional variation in recombination rate, better captures the major determinants of recombination.

It should be noticed that recombination rate can vary during evolution. These variations can be large: for example, the crossover frequency in the telomeric region of the X chromosome is more than ten times lower in *D. melanogaster* than in the closely related species *Drosophila yakuba* (Takano-Shimizu, 2001). These variations can be frequent: for example, several inversions that decrease the local recombination rate are known in the natural populations of *D. melanogaster* (for a review, see Andolfatto *et al*., 2001). MMD01 might reflect better than other methods the average long-term recombination rate along chromosomes. This would explain the stronger correlations with base composition, which is also an average long-term parameter. By contrast, estimates that better capture regional variation in recombination rate should be more strongly correlated with parameters that change over short time-scales. This would explain why, in previous studies, recombination rates have been found to correlate more strongly with nucleotide polymorphism (appearing in the last $4Ne$ generations) than with codon usage (Begun & Aquadro, 1992; Kliman & Hey, 1993; Moriyama & Powell, 1996; for a study of the effects of temporal and spatial variation in recombination rate on genome evolution, see Comeron & Kreitman, 2002). This explanation remains speculative but, if MMD01 was simply more erroneous than other methods, we would have expected to introduce noise into the data and hence to get weaker correlations. Regardless of which method is the most appropriate, it should be stressed that our conclusions remain unchanged when MMD01 is removed from the analysis (see section 3).

(ii) *Recombination and base composition*

We confirm the finding of Marais *et al*. (2001) that the G+C content of large introns and flanking regions is positively correlated with recombination rate, suggesting that neutral substitution patterns vary with recombination in *D. melanogaster*. Additional evidence in this species came from a study by Takano-Shimizu (2001) of substitution patterns in genes located in the telomeric regions of the X chromosomes of several *Drosophila* species. She showed that in *D. melanogaste*r, where the recombination rate is very low in this region, substitutions are AT-biased, for both synonymous sites and non-coding DNA. By contrast, she found that, in *D. yakuba*, in which the recombination rate is more than ten times higher in this region, substitutions are GC-biased in both synonymous sites and non-coding DNA. Such a neutral effect of recombination has also been demonstrated in the nematode (Marais *et al*., 2001), and it has been proposed to occur in a wide range of eukaryotic organisms, such as yeasts (Baudat & Nicolas, 1997; Gerton *et al*., 2000; Birdsell, 2002), mice (Perry & Ashworth, 1999) and humans (Eyre-Walker, 1993; Eisenbarth *et al*., 2000; Yu *et al*., 2001; Fullerton *et al*., 2001; Galtier *et al*., 2001).

Various experimental, sequence analysis and theoretical arguments indicate that this neutral effect of recombination might be due to gene conversion biased towards GC bases (Galtier *et al*., 2001; Birdsell, 2002). The biased gene conversion (BGC) towards GC occurs when the molecular intermediate formed during meiosis contains mismatches. These mismatches are located in heteroduplexes that are DNA stretches with one paternal strand and one maternal strand, and are due to differences in paternal and maternal DNA. It seems that, in general, the repair of such mismatches is biased towards GC: the repair system favours the G/C allele to the A/T allele when the mismatches are A–G, A–C, T–G, or T–C (Birdsell, 2002). The meaning of this bias is not yet fully understood but an expected effect of BGC is the increase of the G+C content of DNA where recombination occurs. It should be realized that the recombination rate has been measured with genetic maps in *D. melanogaster* (see above). Thus, it corresponds to the rate of crossover. Some evidence indicates that the crossover rate and the gene-conversion rate are not totally correlated. For instance, in *D. melanogaster*, chromosome 4 does not make any crossover but does experience some gene conversion (Jensen *et al*., 2002; Wang *et al*., 2002) and the telomeric region of the X chromosome has a very low rate of crossover but a normal rate of gene conversion compared to the remaining of the genome (Langley *et al*., 2000). Thus, when recombination is measured by the rate of crossover, the BGC model does not necessarily predict a strong correlation with base composition. When the recombination rate is measured directly by double-strand-break mapping (which has been done in yeast (Gerton *et al*., 2000)), the correlation between recombination rate and G+C content is much stronger, in agreement with a causal effect of BGC on base composition (Birdsell, 2002).

Our analyses confirm that the association between neutral substitution patterns and recombination rate (possibly explained by BGC) contributes to the positive correlation between codon bias and recombination rate. However, this does not mean that HRi does not exist. As mentioned in our previous article, we think that the neutral and the selective models are not mutually exclusive (Marais *et al*., 2001; Duret, 2002). Indeed, the fact that recombination rate is more strongly correlated with non-coding G+C content than with *Fnop-GC* might reflect such a dual

impact of recombination (both neutral and selective) on synonymous sites. But our analyses show that the impact of HRi on codon usage is obscured by variation in neutral substitution patterns associated with recombination rate. Thus, it is essential to take into account the neutral effect of recombination to be able to detect HRi. Hey & Kliman (2002) attempted to do this by computing the residuals of the regression analysis of *GCnc* and synonymous-codon frequencies. However, *GCnc* might not be the appropriate index to remove precisely the neutral effect of recombination on codon usage, for the reasons already mentioned in section 2 and also because non-coding DNA and synonymous sites might be affected by different neutral substitution patterns (owing to transposable elements, which are often AT-rich and are mainly located in the non-coding DNA (Lerat *et al*., 2000, 2002)). Another possibility would be to analyse genes with low expression level. Such genes have a weak codon bias (Shields *et al*., 1988; Duret & Mouchiroud, 1999) and a high rate of synonymous substitution (Shields *et al*., 1988; Sharp & Li, 1989; Powell & Moriyama, 1997; but see Dunn *et al*., 2001) and so their codon usage is thought to be influenced mainly by neutral substitution pattern.

By using synonymous sites of such lowly expressed genes as markers of neutral substitution patterns, we found that $\sim 4\%$ of the *D. melanogaster* genes have their codon usage affected by HRi and the decrease in codon bias caused by this effect is $\sim 5\%$ (Marais & Piganeau, 2002). This indicates that HRi exists but can be considered a minor determinant of variation in codon bias across the *D. melanogaster* genome. Another piece of evidence from HRi in the *D. melanogaster* genome comes from a recent study by Betancourt and Presgraves (2002), who showed that the interference between advantageous non-synonymous mutations and synonymous mutations could lead to a decrease in codon bias. However, this might only concern genes under positive selection. Although no estimate of the proportion of such genes in the *D. melanogaster* genome is available, it seems reasonable that they are few. Recent theoretical and empirical findings by Comeron & Kreitman (2002) indicate that HRi could explain variation in codon bias inside the *D. melanogaster* genes. However, little is known about the effect of BGC on base composition pattern at the gene scale. Moreover, the current population genetics models of codon usage evolution do not take BGC into account. Their predictions might be changed after incorporating this parameter.

We thank Jody Hey, Antonio Bernardo Carvalho and Josep Comeron for providing their measures of recombination rate in *D. melanogaster*. We are grateful to Nicolas Galtier, Gwenaël Piganeau, Raquel Tavares and both anonymous referees for helpful comments on the manuscript.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

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

Akashi, H., Kliman, R. M. & Eyre-Walker, A. (1998). Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetica* **102/103**, 49–60.

Andolfatto, P., Depaulis, F. & Navarro, A. (2001). Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genetical Research* **77**, 1–8.

Baudat, F. & Nicolas, A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. *Proceedings of the National Academy of Sciences of the USA* **94**, 5213–5218.

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

Betancourt, A. J. & Presgraves, D. C. (2002). Linkage limits the power of natural selection in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA* **99**, 13616–13620.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution* **19**, 1181–1197.

Carvalho, A. B. & Clark, A. G. (1999). Intron size and natural selection. *Nature* **401**, 344.

Comeron, J. M., Kreitman, M. & Aguadé, M. (1999). Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* **151**, 239–249.

Comeron, J. M. & Kreitman, M. (2000). The correlation between intron length and recombination in Drosophila. Dynamic equilibrium between mutational and selective forces. *Genetics* **156**, 1175–1190.

Comeron, J. M. & Kreitman, M. (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**, 389–410.

Dunn, K. A., Bielawski, J. P. & Yang, Z. (2001). Substitution rates in *Drosophila* nuclear genes: implications for translational selection. *Genetics* **157**, 295–305.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics and Development* **12**, 640–649.

Duret, L. & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences of the USA* **96**, 4482–4487.

Eisenbarth, I., Vogel, G., Krone, W., Vogel, W. & Assum, G. (2000). An isochore transition in the *NF1* gene region coincides with a switch in the extent of linkage disequilibrium. *American Journal of Human Genetics* **67**, 873–880.

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings of the Royal Society of London Series B* **252**, 237–243.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* **78**, 737–756.

Fullerton, S. M., Carvalho, A. B. & Clark, A. G. (2001). Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution* **18**, 1139–1142.

Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907–911.

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O. & Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the USA* **97**, 11383–11390.

Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. (1994). Selection intensity for codon bias. *Genetics* **138**, 227–234.

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

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Jensen, M. A., Charlesworth, B. & Kreitman, M. (2002). Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. *Genetics* **160**, 493–507.

Kindahl, E. C. (1994). Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*. Ph.D. Thesis, Cornell University, Ithaca, NY.

Kliman, R. M. & Hey, J. (1993). Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Molecular Biology and Evolution* **10**, 1239–1258.

Kliman, R. M. & Hey, J. (1994). The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137**, 1049–1056.

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

Lerat, E., Biémont, C. & Capy, P. (2000). Codon usage and the origin of *P* elements. *Molecular Biology and Evolution* **17**, 467–468.

Lerat, E., Capy, P. & Biémont, C. (2002). Codon usage by transposable elements and their host genes in five species. *Journal of Molecular Evolution* **54**, 625–637.

Marais, G. & Piganeau, G. (2002). Hill–Robertson interference is a minor determinant of variations in codon bias across *D. melanogaster* and *C. elegans* genomes. *Molecular Biology and Evolution* **19**, 1399–1406.

Marais, G., Mouchiroud, D. & Duret, L. (2001). Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences of the USA* **98**, 5688–5692.

Moriyama, E. N. & Powell, J. R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. *Molecular Biology and Evolution* **13**, 261–277.

Moriyama, E. N. & Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of Molecular Evolution* **45**, 514–523.

Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Research* **20**, 4255–4262.

Perry, J. & Ashworth, A. (1999). Evolutionary rate of a gene affected by chromosomal position. *Current Biology* **9**, 987–989.

Powell, J. R. & Moriyama, E. N. (1997). Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA* **94**, 7784–7790.

Sharp, P. M. & Li, W. H. (1989). On the rate of DNA sequence evolution in *Drosophila*. *Journal of Molecular Evolution* **28**, 398–402.

Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988). 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704–716.

Takano-Shimizu, T. (2001). Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. *Molecular Biology and Evolution* **18**, 606–619.

Wang, W., Thornton, K., Berry, A. & Long, M. (2002). Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science* **295**, 134–137.

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A. J., Deloukas, P., Olsen, A., Doggett, N. A., Ghebranious, N., Broman, K. W. & Weber, J. L. (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953.