Journal of Evolutionary Biology
January 2006, Volume 19 Issue 1, Pages 1 - 11
http://dx.doi.org/10.1111/j.1420-9101.2005.00996.x

Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome

Nicolas Bierne[1, 2, *] and Adam Eyre-Walker[1]

[1] Centre for the Study of Evolution & School of Biological Sciences, University of Sussex, Brighton, BN1 9QG, UK
[2] Laboratoire Génome, Populations, Interactions, Adaptation, UMR5171 CNRS-UMII IFREMER, SMEL, 1 Quai de la daurade, 34200 Sète, France

*: Corresponding author : N. Bierne, email address : n-bierne@univ-montp2.fr

Abstract:

A strong negative correlation between the rate of amino-acid substitution and codon usage bias in Drosophila has been attributed to interference between positive selection at nonsynonymous sites and weak selection on codon usage. To further explore this possibility we have investigated polymorphism and divergence at three kinds of sites: synonymous, nonsynonymous and intronic in relation to codon bias in D. melanogaster and D. simulans. We confirmed that protein evolution is one of the main explicative parameters for interlocus codon bias variation (r^2~ 40% However, intron or synonymous diversities, which could have been expected to be good indicators of local interference [here defined as the additional increase of drift due to selection on tightly linked sites, also called 'genetic draft' by Gillespie (2000)] did not covary significantly with codon bias or with protein evolution. Concurrently, levels of polymorphism were reduced in regions of low recombination rates whereas codon bias was not. Finally, while nonsynonymous diversities were very well correlated between species, neither synonymous nor intron diversities observed in D. melanogaster were correlated with those observed in D. simulans. All together, our results suggest that the selective constraint on the protein is a stable component of gene evolution while local interference is not. The pattern of variation in genetic draft along the genome therefore seems to be instable through evolutionary times and should therefore be considered as a minor determinant of codon bias variance. We argue that selective constraints for optimal codon usage are likely to be correlated with selective constraints on the protein, both between codons within a gene, as previously suggested, and also between genes within a genome.

Keywords: Codon usage, selective constraints, substitution rate, diversity, GC-content, linkage, Hill-Robertson interference, *Drosophila*

Introduction

It is now widely accepted that weak selection for codon usage is acting upon synonymous mutations in some organisms (for review see Kurland, 1991; Sharp *et al.*, 1995; Akashi & Eyre-Walker, 1998) but is relaxed in others (e.g. in some bacteria, Sharp *et al.*, 2005; in Mammals, Duret, 2002). Variation in the intensity of synonymous selection has sometimes been detected at small evolutionary time scales between closely related species: in *Drosophila* for instance, population genetics analyses suggest that selection for codon usage is currently active in *D. simulans* (Akashi & Schaeffer, 1997; Kliman, 1999; Begun, 2001) while it seems to be relaxed in *D. melanogaster* owing to a recent reduction in population size in this lineage (Akashi, 1996; McVean & Vieira, 2001). However, it has proven extremely difficult to distinguish between competing explanations that account for interlocus variance in codon bias within a genome, once selection had been validated at the scale of the whole genome (Akashi, 2001; Duret, 2002). The pattern of codon usage results from a balance among three factors: (i) selection, (ii) random genetic drift and (iii) mutation (Bulmer 1991). All three factors are likely to vary within a recombining genome and have been proposed in turn to be the causative agent.

(i) Interlocus variation in codon usage bias may more simply be the consequence of unequal selection coefficients across genes. The fitness difference between synonymous codons most probably relies on translation efficiency. In *Drosophila*, codon usage is biased toward preferred codons that generally correspond to the most abundant cognate tRNA (Moriyama & Powell, 1997). Variation in translational selection across genes is attested by a positive correlation between codon usage bias and the level of gene expression (Duret & Mouchiroud, 1999). In addition, a variable selective regime on synonymous mutations is further suggested by a negative correlation between codon bias and synonymous substitution rates (Sharp & Li, 1989; Moriyama & Hartl, 1993; Bierne & Eyre-Walker, 2003). However, it

remains unclear whether codon usage primarily affects the elongation rate or the fidelity of protein synthesis (Akashi, 2001; Duret, 2002). The latter hypothesis is supported in *Drosophila* and *Caenorhabditis*, where codon bias is stronger at constrained than at substituted amino acids (Akashi, 1994; Marais & Duret, 2001). However, the translational accuracy hypothesis would predict a positive correlation between codon bias and gene length, as is observed in the Prokaryote *Escherichia coli* (Eyre-Walker, 1996), while the reverse is observed in the metazoan genomes analysed to date (Moriyama & Powell, 1998; Duret & Mouchiroud, 1999; Comeron *et al.*, 1999). How much strengths of selection for the speed or the accuracy of translation vary across genes therefore remains unclear.

(ii) Another factor which can cause within-genome variation in selection efficacy is the Hill-Robertson (HR) effect (Hill & Robertson, 1966). The HR effect corresponds to a decrease in the efficacy of selection acting upon a mutation due to selection on other genetically linked segregating mutations. Selection, whatever its direction, increases the variance in reproductive success and consequently inflates genetic drift (Wright, 1931). The HR effect and related models for neutral mutations (i.e. hitchhiking, Maynard Smith & Haigh, 1974; background selection, Charlesworth *et al.*, 1993) can therefore be understood as local variations in genetic drift (Felsenstein, 1974). Gillespie (2000) suggested that “genetic draft” could prove a useful label for the stochastic effects induced by indirect selection that are different in their origin and their statistic properties from purely demographic random drift. We will here follow Gillespie’s terminology and will state that the magnitude of the genetic draft but not genetic drift can vary along a recombining genome. The correlation observed between local recombination rates and gene diversity in *D. melanogaster* (Begun & Aquadro, 1992) has classically been attributed to higher interferences (be they caused by positive or negative selection) in genomic regions of low recombination, in accordance with this idea of within-genome variation in genetic draft. Although some confounding factors obscure the

correlation (Marais *et al.*, 2001), codon usage bias also correlates with local recombination rates in *D. melanogaster* (Hey & Kliman, 2002; Marais & Piganeau, 2002), as expected by the HR effect. However, the correlation is very weak, accounting for only ~1% of the codon bias variance, and is restricted to lower recombination rate values (Marais & Piganeau, 2002). HR effect, although operating on codon bias (Hey & Kliman, 2002), was therefore thought to be a minor determinant of interlocus codon bias variance (Marais & Piganeau, 2002). However, Betancourt and Presgraves (2002) have documented a very strong negative correlation between the rate of amino-acid substitution and codon usage bias in *Drosophila*. These authors proposed that a small scale HR effect accounts for the correlation: genetic draft would be more intense in fast-evolving genes that undergo a high rate of selectively driven amino-acid substitutions and would thus be unable to optimise their codon usage. The HR effect hypothesis is an indirect interpretation of the correlation that would require further evidence; however Betancourt and Presgraves (2002) discussed and refuted alternative hypotheses making the HR effect the last possible explanation (but see Marais & Charlesworth, 2003). In addition, Kim (2004) has recently shown theoretically that the model is reasonable.

(iii) Finally, codon bias may still vary across genes that share similar selection intensities if a correlated mutational bias is superposed on selection on synonymous codon use. In *Drosophila*, all preferred codons end in G or C which results in a very strong positive correlation ($R^2>0.9$) between codon bias and GC content at synonymous third coding positions (hereafter GC_3). Furthermore, a good correlation ($0.1<R^2<0.3$) is observed between GC_3 and intron GC content (hereafter GC_i) in accordance with the hypothesis that a non negligible mutational bias is superimposed on selection on codon use in this taxa (Kliman & Hey, 1994; Akashi *et al.*, 1998). Marais *et al.* (2001) have pointed out that GC_i correlates positively with local recombination rate, as does GC_3, in *D. melanogaster*, suggesting that a

part of the mutational bias may be associated with recombination. However, the correlation is very weak (R^2<1%), sometimes not detected (Hey & Kliman, 2002), and recombination only accounts for a very small fraction of the correlation between GC_3 and GC_i.

In short, each factor (selection for the speed or the accuracy of translation, genetic draft and mutation bias) appears to be operating concomitantly in *Drosophila*. However, their relative contribution is still unclear. Most importantly, the recent observation that codon usage bias is strongly correlated with the rate of amino-acid substitution, led to the suggestion that small scale variations in local interference within the *Drosophila* genome may have a stronger impact on interlocus codon bias variance than previously thought (Betancourt & Presgraves, 2002; Kim, 2004).

In the present study we propose (i) to explore the effect of within-genome variation in local interference on codon usage bias using a more direct measure of genetic draft, relative levels of synonymous and intron polymorphism, (ii) to use a comparative analysis between *D. melanogaster* and *D. simulans* in order to investigate the relative stability through time of local interference on the one hand, and selective constraint on the protein on the other hand, and (iii) to measure the relative contributions of various factors to synonymous codon bias. Since our sample size is constrained by the availability of polymorphism data at three kinds of sites (intron, synonymous and non-synonymous sites) which could only be found for a few tens of genes at present, we could only study the factors which have a major impact on synonymous codon use.

Data

Dataset 1- Intron and synonymous diversity, synonymous codon bias and within-genome variation in local interference.

One aim of the present work was to investigate intron and synonymous diversity as a correlate of local interference in order to further explore the model of small scale HR effect proposed by Betancourt and Presgraves (2002). To do this, we have compiled from the literature a first dataset composed of 38 genes in *D. melanogaster* and 34 genes in *D. simulans* for which polymorphism data were available at three kinds of sites –intron, synonymous and non-synonymous sites (see Bierne & Eyre-Walker, 2004). Data were available in both species for 23 genes. Unfortunately, the constraint on the data resulted in too few genes sharing a similar sampling scheme in *D. melanogaster* to conduct correlation analyses with reasonable sampling sizes. We therefore have chosen to use as many genes as possible keeping in mind the potential caveat surrounding the dataset. On the other hand, the sampling scheme of the *D. simulans* dataset was very homogeneous as the sequence data comes from a single population in California (Begun & Whitley, 2000). For each locus we have computed the non-synonymous, synonymous and intron diversities within a species as well as substitution rates between species (respectively θc_n, θc_s, θ_i and Dc_n, Dc_s, D_i) using DnaSP (Rozas & Rozas, 1999). Since there is ambiguity surrounding the definition of a site in coding sequences which can lead to a spurious correlation between rates of non-synonymous substitution and codon bias (Muse, 1996; Bierne & Eyre-Walker, 2003), we used synonymous and non-synonymous rates *per codon* instead of *per site*, while intron rates were of course *per site*. As a consequence, intron and synonymous evolutionary rates were not directly comparable; however this was not the purpose of this analysis where a correlative approach was used. In addition, correction for multiple hits could safely be ignored in the polymorphism as well as the divergence between the closely related species *D. melanogaster*

and *D. simulans*. Diversities were estimated from the number of polymorphic sites and sample size (Watterson, 1975) or the average number of nucleotide differences per site between two sequences (Nei, 1987); however the results were very similar with both estimators and we arbitrarily choose Watterson's estimator to present in the results below. Codon usage bias was first measured by the frequency of optimal codons (*Fop*, Ikemura, 1985) by using the program CODONW (Peden, 1999). Other measures of codon bias such as the effective number of codons (ENC, Wright, 1990) or the codon bias index (CBI, Morton, 1993) are sometimes used in the literature. These measures were very well correlated to *Fop* and gave qualitatively similar results. We chose *Fop* because optimal codons have been nicely defined from data on expression levels in *Drosophila* (Duret & Mouchiroud, 1999). As explained above however, *Fop* and other measures of codon bias are correlated to GC content in *Drosophila*, which is taken as evidence for the action of a mutational bias sometimes favouring the same codons as selection does. In order to account for variations in mutational patterns we computed the residuals of the *Fop*/GC_i correlation, that we named *Fop*-GC_i. *Fop*-GC_i is expected to measure the single action of selection freed from the cumulative effect of mutation bias. Finally we estimated recombination rates in *D. melanogaster* (hereafter R_{mel}) by using the data and standard method of Kliman and Hey (1993). We are aware of the debate surrounding the accuracy of different estimates of recombination rates in *Drosophila* (Marais *et al.*, 2003; Kliman & Hey, 2003); however we chose the same estimator as the one used by Betancourt and Presgraves (2002) as a basis for comparison, having verified that other estimates gave qualitatively similar results. The data were compiled into a spreadsheet which is provided in the Supplementary file available on the journal web site.

Dataset 2- relative contribution of each factor

In a second analysis we have investigated the relative contribution of each factor potentially involved in codon bias variation –e.g. protein evolution, expression level, gene length and surrounding non-coding GC content. Since we were not constrained by the need to have intron polymorphism data in this analysis we were able to compile a much larger dataset of genes. Thanks to the recent effort to produce large polymorphism datasets in *D. simulans* (e.g. Begun & Whitley, 2000; Schlenke & Begun, 2003), we were able to compute non-synonymous and synonymous diversities from 105 genes in this species. We chose to measure the level of constraint on a protein using polymorphism data, rather than divergence data, because adaptive substitutions can affect divergence estimates. However, qualitatively similar results were obtained using dN/dS. Codon usage bias was measured by *Fop*. We did not correct for local GC content in this analysis because we were interested in assessing the mutation bias effect (while in the previous dataset we wanted to remove it). Non-coding GC content (hereafter GC_{nc}) usually was GC_i but for the few genes without introns, GC_{nc} was computed from the surrounding non-coding DNA (500 to 1000 bp on either side depending on the distance from adjacent genes) having verified that GC_i and GC_{nc} were very well correlated in the *Drosophila* genome ($R^2=0.99$, $P < 0.001$). Rough estimates of gene expression levels were measured by EST-counting using the procedure described in Duret and Mouchiroud (1999). Finally, gene length was also considered in the analysis as it is known to correlate with codon usage in *Drosophila* (Powell & Moriyama, 1997; Duret & Mouchiroud, 1999). The data were compiled into a spreadsheet which is provided together with dataset 1 in the Supplementary file available on the journal web site.

Analyses

Comparison of the effect of recombination and amino acid substitution rates on silent diversity and codon usage bias in *D. melanogaster*

To begin with, we have explored the possibility that the relevant scale for variations in local interference could be better captured by amino acid substitution rates than by local recombination rates. We were also interested in verifying whether the within-genome variance in local interference was accurately captured by the data in *D. melanogaster* despite an unbalanced sampling scheme between loci. In accordance with a well known observation in *Drosophila* (Begun & Aquadro, 1992), the diversity measured at synonymous and intronic sites (θ_{i+s}) was significantly correlated with recombination (R_{mel}) in our dataset 1 (Fig. 1A). In order to estimate the noise introduced by heterogeneous sampling in our meta-analysis of *D. melanogaster* data, we have plotted, in the same graph, results from a survey which was devoid of bias in the sampling strategy (Andolfatto & Przeworski, 2001). Andolfatto and Przeworski (2001) used sequence data from a single population of Zimbabwe in Africa; few loci, though, were screened simultaneously in exons and introns as required for the present analysis. Figure 1A shows that the correlation we obtained with dataset 1 does not differ greatly from the one obtained by Andolfatto and Przeworski (2001). A Levene's test of homogeneity of variance reveals that the variance of θ are not significantly different in the two datasets ($F_{1,73}$=2.79, P=0.1). In addition, the method of Stephan (1995) was used to fit the curve expected under a model of recurrent selective sweeps (Stephan *et al.*, 1992). The fitted curves were roughly the same with the two datasets (Fig. 1A). This comparison suggests that variations in local interference are accurately captured in dataset 1 and that the heterogeneity of this dataset introduces neither bias nor substantial statistical noise. In the same dataset however, codon usage bias (*Fop*-GC_i) did not correlate significantly with R_{mel} (Fig. 1B) while it was strongly correlated with amino acid substitution rate, Dc_n (Fig. 1D), as previously

reported (Betancourt & Presgraves, 2002; Marais *et al.*, 2004). The hypothesis of a small scale HR effect responsible for low codon bias in fast evolving genes would have predicted a correlation between Dc_n and θ_{i+s}, but this was not the case (Fig. 1C). Conflicting results were therefore obtained depending on whether amino acid substitution rates or local recombination rates were used to assess local interference.

[Figure 1]

The relationship between DNA variation and codon usage bias

Table 1 presents the various correlations obtained between codon bias as measured by *Fop*-GC_i and DNA variation decomposed into three classes of mutations (i.e. non-synonymous, synonymous and introns) within a species (i.e. diversity) and between species (i.e. divergence). In accordance with previous results (Sharp & Li, 1989; Marais *et al.*, 2004), significant negative correlations were obtained between codon bias and non-synonymous, as well as synonymous substitution rates. In contrast, intron divergence did not correlate significantly with codon bias (Table 1), or with GC_i (r_s=-0.12, n.s.), in accordance with the neutral expectation. The correlation previously observed with non-synonymous divergence was here extended to non-synonymous diversity within both species. Non-synonymous diversity is most likely composed of neutral or nearly neutral mutations and it was unclear whether one would have expected non-synonymous diversity to be a good index of the density of selected mutations and local genetic draft. Nevertheless, non-synonymous diversity is known to be highly correlated with non-synonymous divergence in *Drosophila* because selective constraint on the protein is the main determinant of non-synonymous variation (see Bierne and Eyre-Walker 2004). Therefore, we would not take this observation as an argument against the HR effect hypothesis. Synonymous diversity, on the other hand, was not significantly correlated with codon bias which was predictable in *D. melanogaster* where

synonymous selection is thought to be relaxed (Akashi, 1996) but could have been expected in *D. simulans*. Indeed, the results obtained with synonymous mutations were consistent with the effect of weak selection acting on synonymous codons (Bulmer, 1991). Finally, and most importantly, intron diversity, which could have been expected to be an unbiased indicator of local interference, did not covary significantly with codon bias (Table 1).

[Table 1]

Comparative analysis

Our dataset 1 allowed us to compare diversities between different classes of mutations within a species, and for the same class of mutations, compare the diversity realised in each species. Such an analysis was conducted to further investigate how local interference could evolve between species. Both in *D. melanogaster* and *D. simulans*, synonymous and intron diversities were strongly correlated (Fig. 2A). However, neither intron nor synonymous diversity within *D. melanogaster* was correlated with intron or synonymous diversity in *D. simulans* (Fig. 2B). Non-synonymous diversity on the other hand did not correlate significantly with intron (Fig. 2C) or synonymous (not shown, *D. melanogaster*: r_s=0.29, n.s.; *D. simulans*: r_s=0.13, n.s.) diversity. Non-synonymous diversity within *D. melanogaster* however was strongly correlated with non-synonymous diversity within *D. simulans* (Fig. 2D).

[Figure 2]

Taken together, these results illustrate that intron and synonymous variations are mainly driven by stochastic processes (genetic drift and draft) that are not stable components through evolutionary times, while non-synonymous variation is mainly driven by selective constraint on the protein which in contrast seems to be a stable element, at least between closely related species.

The relative contribution of each factor

We can now reconsider the relative contribution of various factors thought to be involved in the codon bias variance with our dataset 2. Correlation statistics are presented in Table 2. We chose the ratio $\omega=\theta c_n/\theta c_s$ as a measure of the selective constraint on the protein (small ω indicates more constraint on the amino acid sequence). In decreasing order the parameters that appeared to explain most of the variation in codon usage were (i) the selective constraint on the protein as measured by the ω ratio, which explains ~40% of the codon bias variance, (ii) the local mutational pattern as measured by surrounding non-coding GC content which explains ~15% of the codon bias variance and (iii) expression levels as measured by EST-counting which explains ~10% of the codon bias variance. None of these three parameters co-vary significantly with each other in this dataset suggesting that they correspond to almost independent factors. Note however that a correlation between the level of gene expression and non-synonymous evolutionary rates has been described elsewhere, although with a much larger dataset (Marais *et al.*, 2004) or in other organisms, where selection on codon usage is relaxed such as Mammals (Duret & Mouchiroud, 2000). Finally, the correlation between codon bias and gene length which has previously been reported with very large datasets (Powell & Moriyama, 1997; Duret & Mouchiroud, 1999) was not significant in our dataset 2 (Table 2).

[Table 2]

Discussion

Polymorphism and divergence data at three kinds of sites –synonymous, nonsynonymous and intronic– were used to investigate the importance of within-genome variations in local interference on the evolution of codon usage in *Drosophila*. We first argue that our results suggest that fast evolving genes do not have conspicuously higher levels of genetic draft. In addition, a comparative analysis between *D. melanogaster* and *D. simulans* suggests that local interference is unlikely to be a stable component of gene evolution while selective constraint on the protein is. All together our results suggest that the correlation between synonymous codon usage and protein evolution cannot be exclusively interpreted by local interference between selection at non-synonymous and synonymous sites. We will finally discuss alternative explanations involving some connections between selection on the protein and selection for the speed or the accuracy of translation.

Synonymous and intron diversities do not corroborate a more intense genetic draft in the recent history of fast-evolving genes

Since the publication of the correlation between local recombination rates and gene diversity (Begun & Aquadro, 1992), local variation in genetic draft within the *Drosophila* genome has been thoroughly investigated. Using recombination rates to assess the intensity of genetic draft, within-genome variation in local interference has been suspected to influence the efficacy of weak selection on various genomic components such as intron length (Carvalho & Clark, 1999) or codon usage (Kliman & Hey, 1993; Comeron *et al.*, 1999). More recently, other parameters thought to correlate with local interference have been investigated such as gene length, the presence/absence of introns, or the spatial situation of targeting sites in the gene (Comeron & Kreitman, 2002). Most of these correlations are minute, accounting for a minor part of the total variance and thus require very large datasets (often exhaustive genome-

wide datasets) to be detected. In addition, some confounding mutational biases have sometimes been identified (Marais *et al.*, 2001). The weakness of the correlation are perhaps not surprising given the relevant estimates of recombination and mutation rates (Marais & Piganeau, 2002). On the contrary, the correlation observed between the rate of protein evolution and codon bias is surprisingly strong (R^2>40%). It is so strong that it does not require a large dataset to detect; neither does it require the presence of genes with particularly high rates of amino-acid substitution. As a consequence, if the correlation was entirely due to HR effects, one could have expected a detectable effect on levels of polymorphism (McVean & Charlesworth, 2000). However, neither synonymous nor intron diversities significantly correlate with protein evolution nor do they correlate with codon bias. In the same dataset, diversities were significantly reduced in region of low recombination rates whereas codon bias was not. Therefore, it seems difficult to summarise the results obtained within a single framework, namely HR effects.

Drosophila populations are known to exhibit complex patterns of genetic diversity that are not consistent with any simple model at demographic equilibrium (Andolfatto & Przeworski, 2000; Begun, 2001; Wall et al., 2002). *D. melanogaster* and *D. simulans* are thought to have spread across the world from Africa after the last glaciation (David & Capy, 1988). Derived populations are known to depart from demographic equilibrium (Begun & Aquadro, 1993; Begun, 2001; Baudry *et al.*, 2004) but the situation in Africa is not straightforward either (Glinka *et al.*, 2003; Veuille *et al.*, 2004). Indeed, it is likely that natural populations never conform to the standard population genetic assumptions (Lewontin, 2002). One may therefore suspect that departures from equilibrium could introduce unpredicted stochastic variance preventing any solid interpretation of the data. However, we would argue that (i) equilibrium does not need to be assumed here as demographic processes should affect the whole genome in a similar way such that within-genome variation captured in a correlation analysis can only

come from non-demographic processes (*i.e.* genetic draft), (ii) the significant correlation obtained between polymorphism levels and recombination rates attests that a fraction of the within-genome variation in local interference is accurately captured in the data and (iii) for a factor to have a bearing on the long term evolution of a trait with such a minuscule phenotypic consequence as codon bias, its effect should probably surpass the stochastic variance inevitably generated in every natural population. In our dataset, two correlations –between recombination and silent diversity and between non-synonymous polymorphism and codon bias– have proved to persist despite enduring the stochastic pressure.

Alone, though, the apparently conflicting observations we reported are not sufficient to completely refute small scale HR effects because codon bias, polymorphism levels and recombination are parameters that evolve at different time-scales. Diversity may not be reduced in fast-evolving genes nowadays but might have been in the past. Because codon bias depends on long-term evolution (Marais *et al.*, 2004), forces acting on it should be rather stable components of gene evolution.

Local interference is not a stable component of gene evolution

Local interference depends on the density of selected sites, the strength of the selection acting on selected sites and local recombination rates (McVean & Charlesworth, 2000; Stephan & Kim, 2002). Evidence has recently accumulated which suggests that local recombination rates are not stable over even short timescales (e.g. Munte *et al.*, 2001; Takano-Shimizu, 2001; Meunier & Duret, 2004). For instance, Ptak *et al.* (2005) have demonstrated that the recombination landscape has markedly changed during the human/chimp divergence. These results would suggest that local interference may vary accordingly in time. However, the possibility remains that the variation in local selection (density and strength of selection) prevails over the variation in local recombination rate, as implicitly assumed in the model of

Betancourt and Presgraves (2002). To assess the stability of local interference, we have here conducted a comparative analysis of polymorphism levels between *D. melanogaster* and *D. simulans*. Neither intron nor synonymous diversity within *D. melanogaster* was correlated with intron or synonymous diversity in *D. simulans*, suggesting that local interference is not a very stable component of gene evolution. Instead of the apparent stochastic nature of silent diversity, non-synonymous diversities were very well correlated between species suggesting that selective constraints are conserved across species. In accordance with this view, Munte *et al.* (2001) showed that the recombinational environment of a gene strongly conditions synonymous substitution rates while it has no detectable effect on amino acid evolutionary rates in *Drosophila*.

All together, our results suggest that the selective constraint on the protein is a stable component of gene evolution (also see Skibinski & Ward, 2004) while local interference is not.

Correlated selective constraints on synonymous and non-synonymous sites

Our evidence suggests that HR effects are not a strong determinant of codon bias, but why then is there a correlation between synonymous codon usage and rates of protein evolution? The alternatives have been well discussed elsewhere (Akashi, 1994; Betancourt & Presgraves, 2002; Marais *et al.*, 2004) but we reiterate them here briefly. Although attractive at first sight, non-synonymous changes that transform a preferred codon into an unpreferred codon (Lipman & Wilbur, 1985) cannot reasonably account for the correlation. Indeed, removing such codons (which represent 19% of nonsynonymous changes) has no effect on the strength of the correlation between the rate of non-synonymous substitution and codon usage bias (data not shown, see Akashi, 1994; Marais & Duret, 2001; Betancourt & Presgraves, 2002).

It is also easy to refute another possibility –that the correlation arises through the way in which sites are counted in the estimation of the non-synonymous substitution rate. In the method of Goldman and Yang (1994), the method used by Betancourt and Presgraves (2002), sites are counted as mutational opportunities (see Bierne and Eyre-Walker 2003), so as codon bias increases the number of synonymous sites decreases and the number of non-synonymous sites increases. This means that genes with high codon bias will tend to have lower rates of non-synonymous substitution per site (i.e. if two genes have undergone similar numbers of non-synonymous substitutions per codon, the gene with the higher level of codon bias will actually have a lower rate of non-synonymous substitution per site). However, the rate of nonsynonymous substitution *per codon* is also correlated to codon bias.

This leaves an idea originally proposed by Akashi (1994), that the strength of selection acting upon synonymous mutations is correlated to that acting upon non-synonymous mutations. This could be due to selection on translational accuracy –genes in which most amino acid sites need to be occupied by a particular amino acid will evolve slowly and will need to accurately translate. Betancourt and Presgraves (2002) offered several lines of evidence against this hypothesis. First they noted that the rate of synonymous substitution was positively correlated to codon bias in their analysis while it was generally accepted the correlation was negative. However, this was an artefact of the method they used, as we have discussed elsewhere (Bierne & Eyre-Walker, 2003) –the rate of synonymous substitution does correlate negatively and significantly with codon bias (Table 1) as previously reported (Sharp & Li, 1989). Second, Betancourt and Presgraves (2002) tested this hypothesis by considering the correlation between the level of codon bias in codons which had not undergone a non-synonymous substitution and the overall rate of non-synonymous substitution. They found the correlation was unchanged and concluded that there was no evidence of correlated strengths of selection. To explain the logic of their test let us consider a pair of two fold degenerate

codons - phenylalanine for example. Let us imagine that the average strength of selection against non-synonymous mutations is s_n. Errors during translation will have an effect on the fitness of the individual which is correlated to this average strength (the correlation will not be perfect, because while TTT to TTA mutations might be common, TTT to TTA translational errors may not be). This will manifest itself as selection on synonymous codon bias; so the strength of selection on codon bias will be correlated to the strength of selection against deleterious mutations. The average strength of selection against non-synonymous mutations s_s is therefore equal to $k\ s_n$, where k is a constant. It seems likely, unless the translational error rate is very high that $k<1$. Let us now think about all the phenylalanines in a gene. Some will be very important because they are critical for function and others will not be. We can divide the sites into three categories; (i) sites at which $N_e s_s<1$ and $N_e s_n<1$ – i.e. selection at both sites is ineffective; (ii) sites at which $N_e s_s<1$ and $N_e s_n>1$ – sites at which selection is effective against the non-synonymous mutations, but ineffective on synonymous codon use; and (iii) sites at which $N_e s_s>1$ and $N_e s_n>1$ – codons at which selection is effective on both non-synonymous and synonymous mutations. The rate of non-synonymous substitution, ignoring adaptive evolution is determined by the proportion of sites in category (i) relative to categories (ii) and (iii), while the level of synonymous codon use is determined by the proportion of sites in (i) and (ii) relative to (iii). Betancourt and Presgraves (2002) just looked at synonymous codon use at codons with no amino acid substitution which would be equivalent to looking at the relative number of codons in category (ii) versus (iii). It is clear that if the s_n's in a gene are independently and randomly drawn from some distribution then there will be no correlation between the rate of nonsynonymous substitution and the level of bias in codons which have not undergone amino acid substitution – this would be equivalent to randomly allocating codons to the three categories and so there is no expectation of a correlation between (i)/(i+ii+iii) and ii/(ii+iii). However, their test is not valid if there is a

correlation between s_n at different sites within a gene; *i.e.* if genes with strong selection against non-synonymous mutations at one codon also tend to have strong selection at other codons. This is indeed the case – for example the two halves of a gene have correlated rates of non-synonymous substitution (Smith and Eyre-Walker 2002).

The strong correlation between codon bias and rates of non-synonymous substitution, or levels of non-synonymous polymorphism, and our explanation for the correlation, suggest that selection on codon usage bias is primarily driven by translational accuracy. This is supported by the fact that constrained codons tend to have higher levels of codon bias (Akashi, 1994; Marais & Duret, 2001). However, this effect was not very strong and the positive correlation between codon bias and gene length predicted by the translational accuracy hypothesis (Eyre-Walker, 1996) was negative instead (Duret & Mouchiroud, 1999). Codon bias is expected to be stronger in longer genes under translational accuracy because mistakes in longer genes will be energetically more costly. However, controlling for gene function is difficult in this type of analysis – i.e. it may be that longer genes tend to be poorly constrained and therefore fast evolving. If the inter-locus variance in selection regime overwhelms the gene length effect, the correlation would not necessarily be found. Furthermore, the fact that the correlation was not strong could be explained by an overabundance of non-selectively constrained codons in the non-substituted class when the comparison involved closely related species (Akashi, 1994) and an overabundance of selectively constrained codons in the substituted class (i.e. covarion-like evolution, Fitch, 1971) when the comparison involved distant species (Marais & Duret, 2001). Finally, the variance in selective constraints for optimal protein synthesis may be more easily encapsulated between genes within a genome than it is between codons within a gene. Indeed, it is likely that very constrained proteins that play a major role in the correlation, are constrained at nearly every amino-acid, the reverse being true for fast-evolving proteins.

Conclusion

We found evidence against HR effects as a suitable explanation for the correlation between the rate of amino-acid substitution and codon usage bias in *Drosophila*. Although there are theoretical reasons to believe (Hill & Robertson, 1966; McVean & Charlesworth, 2000; Kim, 2004) and empirical data to suggest (Hey and Kliman 2002) that HR interferences are operating on codon bias they cannot reasonably explain such a strong correlation and should be viewed as a minor determinant of interlocus codon bias variance (Marais and Piganeau 2002).

We would therefore conclude that variation in codon usage within the *Drosophila* genome is mainly a simple consequence of unequal selection coefficients across genes. Discriminating between selection for the speed or the accuracy of protein synthesis is difficult but our analysis suggests that the fidelity of translation may be a more important component than previously thought. Usually investigated at the codon level within a gene, the effect of selection on the accuracy of translation may more markedly be seen at the gene level within a genome.

Acknowledgements

We are very grateful to people of the Centre for the Study of Evolution for helpful discussions on the issue of within-genome variations in effective population size and to two anonymous referees for insightful comments on the manuscript. The authors were supported by the Biotechnology and Biological Sciences Research Council and the Royal Society.

References

Akashi, H. 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927-935.

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. 2001 Gene expression and molecular evolution. Curr. Op. Genet. Dev. **11:** 660-666.

Akashi, H., and A. Eyre-Walker 1998 Translational selection and molecular evolution. Curr. Op. Genet. Dev. **8:** 688-693.

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

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

Andolfatto, P., and M. Przeworski 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257-268.

Andolfatto, P., and M. Przeworski 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657-665.

Baudry, E., B. Viginier and M. Veuille 2004 Non-African populations of Drosophila melanogaster have a unique origin. Mol Biol Evol **21:** 1482-1491.

Begun, D. 2001 The frequency distribution of nucleotide variation in *Drosophila simulans*. Mol. Biol. Evol. **18:** 1343-1352.

Begun, D., and C. F. Aquadro 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548-550.

Begun, D. J., and C. F. Aquadro 1992 Levels of naturally occuring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

Begun, D. J., and P. Whitley 2000 Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proceedings of the National Academy of Sciences of the United States of America **97:** 5960-5965.

Betancourt, A. J., and D. C. Presgraves 2002 Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci U S A **99:** 13616-13620.

Bierne, N., and A. Eyre-Walker 2003 The problem of counting sites in the estimation of the synonymous and non-synonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics **165:** 1587–1597.

Bierne, N., and A. Eyre-Walker 2004 The genomic rate of adaptive amino-acid substitution in *Drosophila*. Mol Biol Evol **21:** 1350-1360.

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

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

Charlesworth, B., M. T. Morgan and D. Charlesworth 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

Comeron, J. M., and M. Kreitman 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389-410.

Comeron, J. M., M. Kreitman and M. Aguade 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151:** 239-249.

David, J. R., and P. Capy 1988 Genetic variation of Drosophila melanogaster natural populations. Trends Genet. **4:** 106-111.

Duret, L. 2002 Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12:** 640-649.

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

Duret, L., and D. Mouchiroud 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17:** 68-74.

Eyre-Walker, A. 1996 Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy? Mol. Biol. Evol. **13:** 864-872.

Felsenstein, J. 1974 The evolutionary advantage of recombination. Genetics **78:** 737-756.

Fitch, W. M. 1971 Rate of change of concomitantly variable codons. J Mol Evol **1:** 84-96.

Gillespie, J. H. 2000 Genetic drift in an infinite population: The pseudohitchhiking model. Genetics **155:** 909-919.

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

Goldman, N., and Z. Yang 1994 A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725-736.

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

Hill, W. G., and A. Robertson 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8:** 269-294.

Ikemura, T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13-34.

Kim, Y. 2004 Effect of strong directional selection on weakly selected mutations at linked sites: Implication for synonymous codon usage. Mol Biol Evol **21:** 286 - 294.

Kliman, R. 1999 Recent selection on synonymous codon usage in *Drosophila*. J. Mol. Evol. **49:** 343-351.

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

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

Kliman, R. M., and J. Hey 2003 Hill-Robertson interference in Drosophila melanogaster: reply to Marais, Mouchiroud and Duret. Genetical Research **81:** 89-90.

Kurland, C. G. 1991 Codon bias and gene expression. FEBS Letters **285:** 165-169.

Lewontin, R. C. 2002 Directions in evolutionary biology. Annu Rev Genet **36:** 1-18.

Lipman, D. J., and W. J. Wilbur 1985 Interaction of silent and replacement changes in eukaryotic coding sequences. J. Mol. Evol. **21:** 161-167.

Marais, G., and B. Charlesworth 2003 Genome evolution: recombination speeds up adaptive evolution. Curr. Biol. **13:** R68-R70.

Marais, G., T. Domazet-Loso, D. Tautz and B. Charlesworth 2004 Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. J Mol Evol **59:** 771-779.

Marais, G., and L. Duret 2001 Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. **52:** 275-280.

Marais, G., D. Mouchiroud and L. Duret 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A **98:** 5688–5692.

Marais, G., D. Mouchiroud and L. Duret 2003 Neutral effect of recombination on base composition in Drosophila. Genetical Research **81:** 79-87.

Marais, G., and G. Piganeau 2002 Hill-Robertson interference is a minor determinant of variation in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genome. Mol. Biol. Evol. **19:** 1399-1406.
Maynard Smith, J., and J. Haigh 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.
McVean, G., and J. Vieira 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245-257.
McVean, G. A., and B. Charlesworth 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929-944.
Meunier, J., and L. Duret 2004 Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol **21:** 984-990.
Moriyama, E. N., and D. L. Hartl 1993 Codon usage bias and base composition of nuclear genes of *Drosophila*. Genetics **134:** 847-858.
Moriyama, E. N., and J. R. Powell 1997 Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. **45:** 514-523.
Moriyama, E. N., and J. R. Powell 1998 Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucl. Acids Res. **26:** 3188-3193.
Morton, B. R. 1993 Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. J Mol Evol **37:** 273-280.
Munte, A., M. Aguade and C. Segarra 2001 Changes in the recombinational environment affect divergence in the yellow gene of *Drosophila*. Mol Biol Evol **18:** 1045-1056.
Muse, S. V. 1996 Estimating the synonymous and nonsynonymous substitution rates. Mol Biol Evol **13:** 105-114.
Nei, M. 1987 *Molecular evolutionary genetics*. Columbia University Press, New York.
Peden, J. F. 1999 Analysis of codon usage, pp. University of Nottingham, UK.
Powell, J. R., and E. N. Moriyama 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784-7790.
Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil *et al.* 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet. **18:** [Epub ahead of print].
Rozas, J., and R. Rozas 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.
Schlenke, T. A., and D. J. Begun 2003 Natural selection drives drosophila immune system evolution. Genetics **164:** 1471-1480.
Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi and J. F. Peden 1995 DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci **349:** 241-247.
Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett 2005 Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. **33:** 1141-1153.
Sharp, P. M., and W.-H. Li 1989 On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28:** 3398-3402.
Skibinski, D. O., and R. D. Ward 2004 Average allozyme heterozygosity in vertebrates correlates with Ka/Ks measured in the human-mouse lineage. Mol Biol Evol **21:** 1753-1759.
Stephan, W. 1995 An improved method for estimating the rate of fixation of favourable mutations based on DNA polymorphism data. Mol. Biol. Evol. **12:** 959-962.
Stephan, W., and Y. Kim 2002 Recent applications of diffusion theory to population genetics, pp. 72-93 in *Modern developments in theoritical population genetics, the legacy of*

Gustave Malecot, edited by M. Slatkin and M. Veuille. Oxford University Press, Oxford.

Stephan, W., T. H. E. Wiehe and M. W. Lenz 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor Pop Biol **41:** 237-254.

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

Veuille, M., E. Baudry, M. Cobb, N. Derome and E. Gravot 2004 Historicity and the population genetics of *Drosophila melanogaster* and *D. simulans*. Genetica **120:** 61-70.

Wall, J. D., P. Andolfatto and M. Przeworski 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203-216.

Watterson, G. A. 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256-276.

Wright, F. 1990 The 'effective number of codons' used in a gene. Gene **87:** 23-29.

Wright, S. 1931 Evolution in Mendelian populations. Genetics **16:** 97-159.

Table 1. Spearman's correlation coefficients between codon bias (*Fop*-GC_i) and DNA variation, and significance levels. Significant correlations are in bold. In brackets are the results obtained with the larger dataset, dataset 2, where possible.

	Fop-GC_i	Dc_n
Non-synonymous		
Dc_n	**-0.63*****	
θc_n sim	**-0.71***** **(-0.51*******)**	**0.69***** **(0.62*******)**
θc_n mel	**-0.36***	**0.52*****
Synonymous		
Dc_s	**-0.42****	**0.41****
θc_s sim	-0.27^{NS} (0.08^{NS})	-0.11^{NS} (-0.09^{NS})
θc_s mel	0.20^{NS}	0.20^{NS}
Intron		
D_i	-0.10^{NS}	0.00^{NS}
θ_i sim	0.02^{NS}	0.18^{NS}
θ_i mel	0.18^{NS}	-0.17^{NS}

^{NS}: not significant; * 0.01<p<0.05; **: 0.001<p<0.01; ***:p<0.001.

Table 2. Spearman's correlation coefficients between codon bias (*Fop*) and explicative parameters and significance levels (see the data section). Significant correlations that remained unchanged after a partial correlation analysis are in bold.

	Fop	GC_{nc}	Ln(#ESTs)	Gene length
ω	**-0.65*****	-0.30[a]	-0.07^{NS}	-0.01^{NS}
GC_{nc}	**0.44*****		0.12^{NS}	-0.01^{NS}
Ln(#ESTs)	**0.25****			-0.01^{NS}
Gene length	0.00^{NS}			

^{NS}: not significant; * 0.01<p<0.05; **: 0.001<p<0.01; ***:p<0.001;
[a] not significant after a partial correlation analysis (r_p=-0.05, n.s.).

Figure legends

Figure 1. Comparison of the effect of recombination and the rate of amino-acid substitution on silent diversity and codon usage bias in *Drosophila melanogaster*. (A) Scatterplot of recombination (R_{mel}) and diversity measured at synonymous and intronic sites (θ_{i+s}) in our dataset 1 (black dots) or silent diversity reported in Andolfatto and Przeworski (2001) (crosses). The solid and dotted lines are the best fit of Stephan's model (1995) to dataset 1 and data from Andolfatto and Przeworski (2001), respectively. (B) Scatterplot of recombination (R_{mel}) and codon bias calibrated by intron GC-content (*Fop*-GC_i). (C) Scatterplot of nonsynonymous substitution rate (Dc_n) and diversity (θ_{i+s}). (D) Scatterplot of nonsynonymous substitution rate (Dc_n) and codon bias (*Fop*-GC_i).

Figure 2. Comparison of diversity between different classes of mutations within a species, and comparison of diversity between species for the same class of mutations. (A) Scatterplot of intron diversity (θ_i) and synonymous diversity (θc_s) in *D. simulans* (black dots) and *D. melanogaster* (open dots). (B) Scatterplot of diversity in *D. simulans* and *D. melanogaster* at synonymous sites (black dots) and intronic sites (open dots). (C) Scatterplot of intron diversity (θ_i) and nonsynonymous diversity (θc_n) in *D. simulans* (black dots) and *D. melanogaster* (open dots). (D) Scatterplot of diversity in *D. simulans* and *D. melanogaster* at nonsynonymous sites.

Figure 1

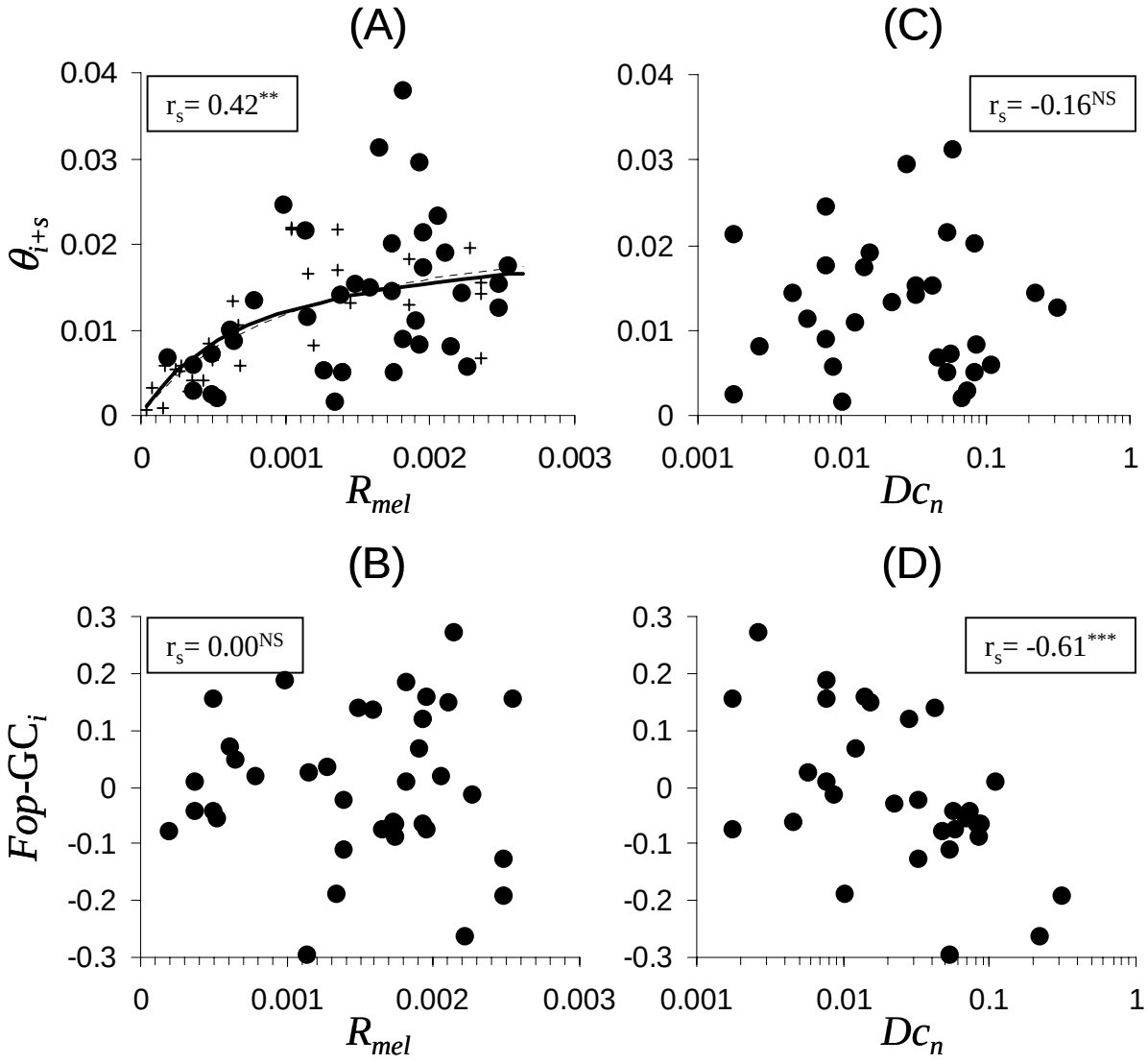
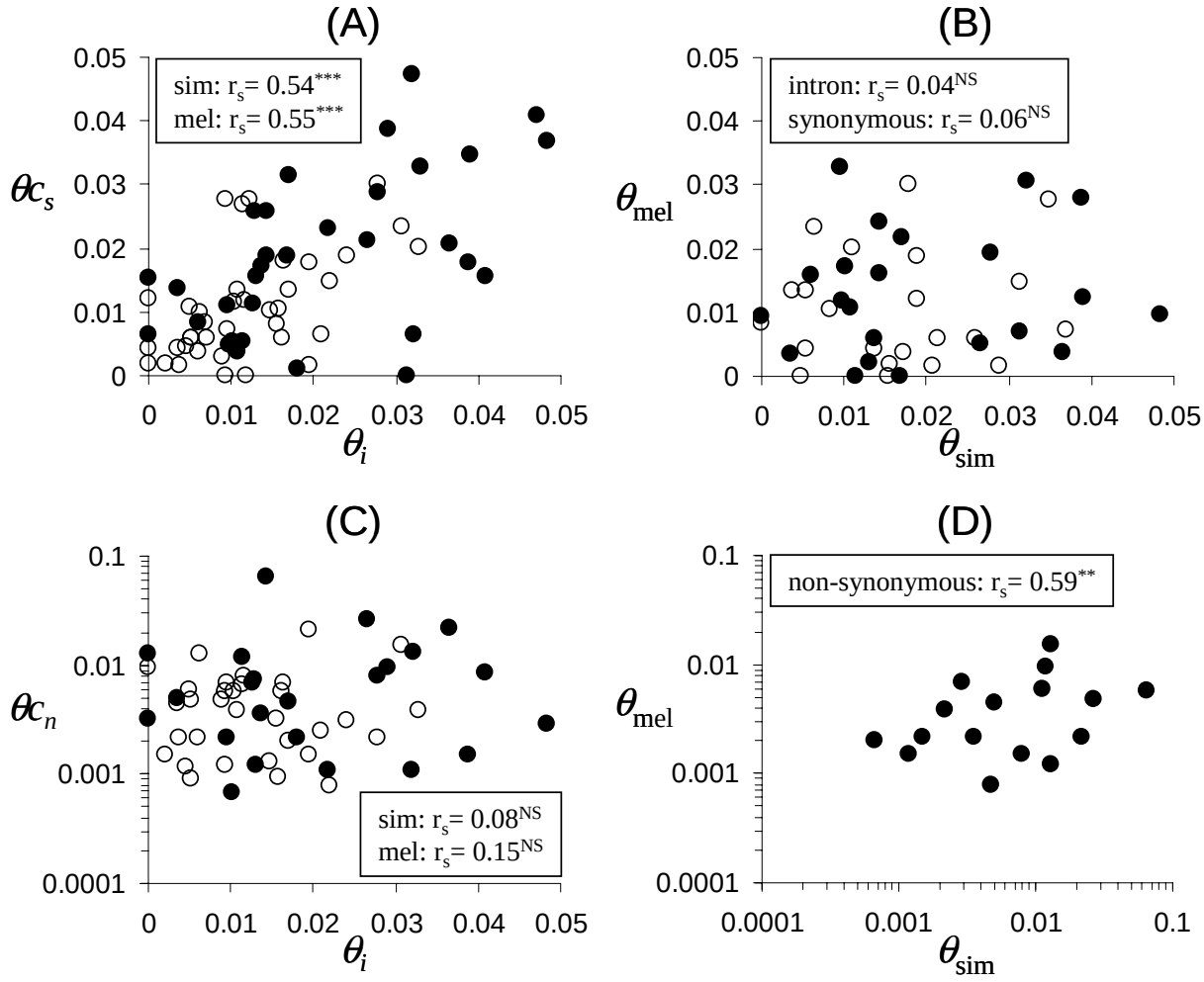

(A)
(C)
(B)
(D)
r_s= 0.42**
r_s= -0.16NS
r_s= 0.00NS
r_s= -0.61***
θ_{i+s}
Fop-GC$_i$
R_{mel}
Dc_n
0.04
0.03
0.02
0.01
0
0.3
0.2
0.1
-0.1
-0.2
-0.3
0.001
0.002
0.003
0.01
0.1
1

Figure 2


(A)
sim: r_s= 0.54***
mel: r_s= 0.55***
θc_s
θ_i
(B)
intron: r_s= 0.04NS
synonymous: r_s= 0.06NS
θ_{mel}
θ_{sim}
(C)
sim: r_s= 0.08NS
mel: r_s= 0.15NS
θc_n
θ_i
(D)
non-synonymous: r_s= 0.59**
θ_{mel}
θ_{sim}