Late-Replicating Domains Have Higher Divergence and Diversity in *Drosophila melanogaster*

Claudia C. Weber, Catherine J. Pink, and Laurence D. Hurst*

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

***Corresponding author:** E-mail: l.d.hurst@bath.ac.uk.

Associate editor: John Parsch

Abstract

Several reports from mammals indicate that an increase in the mutation rate in late-replicating regions may, in part, be responsible for the observed genomic heterogeneity in neutral substitution rates and levels of diversity, although the mechanisms for this remain poorly understood. Recent evidence also suggests that late replication is associated with high mutability in yeast. This then raises the question as to whether a similar effect is operating across all eukaryotes. Limited evidence from one chromosome arm in *Drosophila melanogaster* suggests the opposite pattern, with regions overlapping early-firing origins showing increased levels of diversity and divergence. Given the availability of genome-wide replication timing profiles for *D. melanogaster*, we now return to this issue. Consistent with what is seen in other taxa, we find that divergence at synonymous sites in exon cores, as well as divergence at putatively unconstrained intronic sites, is elevated in late-replicating regions. Analysis of genes with low codon usage bias suggests a ~30% difference in mutation rate between the earliest and the latest replicating sequence. Intronic sequence suggests a more modest difference. We additionally show that an increase in diversity in late-replicating sequences is not owing to replication timing covarying with the local recombination rate. If anything, the effects of recombination mask the impact of replication timing. We conclude that, contrary to prior reports and consistent with what is seen in mammals and yeast, there is indeed a relationship between rates of nucleotide divergence and diversity and replication timing that is consistent with an increase in the mutation rate during late S-phase in *D. melanogaster*. It is therefore plausible that such an effect might be common among eukaryotes. The result may have implications for the inference of positive selection.

Key words: replication timing, mutation rate, divergence, diversity, codon usage bias.

Introduction

It is well established that in mammalian genomes, neighboring genes show more similar rates of evolution at intronic or exonic synonymous sites than would be expected by chance (see e.g., Wolfe et al. 1989; Matassi et al. 1999; Smith et al. 2002; Lercher et al. 2004). The reasons for this local similarity (regional homogeneity) are not well characterized. Recent reports have implicated replication timing, with sequences replicating late during S-phase exhibiting increased neutral divergence and diversity due to what appears to be an increase in the mutation rate (Stamatoyannopoulos et al. 2009; Chen et al. 2010; Pink and Hurst 2010). While the mechanisms responsible for this remain to be elucidated (see Discussion in Chen et al. 2010), a similar effect has been suggested to occur in *Saccharomyces cerevisiae* (Lang and Murray 2011).

That such phylogenetically distant species as yeast and mammals both show evidence of higher mutation rates for late-replicating sequence tempts the speculation that this might be a universal eukaryotic feature. However, an increase in both heterozygosity within *Drosophila simulans* and divergence between *D. simulans* and *D. melanogaster* has been reported near early-firing replication origins (as determined in *D. melanogaster* Kc cells, see MacAlpine et al. 2004) on chromosome 2L (Begun et al. 2007). This suggests that *Drosophila* may show the opposite trend to that seen in yeast and mammals. Moreover, while polymorphism and divergence are known to show large-scale fluctuations in *D. simulans*, this has been attributed to the effects of natural selection coupled with differences in the recombination rate rather than heterogeneous mutation rates (Begun et al. 2007). It is also not known whether neighboring genes have similar rates of evolution at putatively neutral sites as observed in mammals.

Given several differences in genome organization between mammals and drosophilids, it is perhaps not obvious that we should necessarily expect to observe the same trends in the two groups. For example, while the increase in CpG substitutions in late-replicating mammalian sequences has been attributed to methylation levels rising during S-phase (Chen et al. 2010), *D. melanogaster* lacks nuclear methylation (measured in 0–3 h embryos) (Zemach et al. 2010). Moreover, in mammals, isochore boundaries correspond to boundaries between late- and early-replicating sequence (Schmegner et al. 2005, 2007). In contrast, no significant difference in GC content between cytologically defined late-replicating regions and other regions has been reported in *D. melanogaster* (Belyakin et al. 2010). Replication timing zones in *D. melanogaster* do, however, have a number of other genomic features. Lamin B targets, which associate with the nuclear periphery, are late replicating, as well as being more likely to be found in regions of conserved gene order, lacking active histone marks and being transcriptionally silent (Pickersgill et al. 2006; Ranz et al. 2011). Clusters of testis-expressed genes also tend to be found in late-replicating regions

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Research article

http://mbe.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

(Belyakin et al. 2005). Similar associations between replication timing and subnuclear localization and chromatin structure have been described in mammals (see e.g., Williams et al. 2006; Grasser et al. 2008; Hiratani et al. 2008, 2010). Additionally, late-replicating sequences in yeast show a similar pattern of association with the nuclear periphery, as well as being hypoacetylated and heterochromatic, indicating that these features may be shared between all eukaryotes (see Gilbert 2002; Donaldson 2005).

Given the recent availability of genome-wide replication timing data for *D. melanogaster* (Schwaiger et al. 2009), we now provide the first genome-scale assessment of the relationship between replication timing and divergence and diversity at putatively neutral sites. We first ask whether a local similarity in synonymous substitution rates exists in flies. We then attempt to establish whether replication timing and mutation might be related. Consistent with what is seen in other taxa, we find that both divergence and diversity are elevated in late-replicating regions. We attempt to estimate the magnitude of the effect by considering what happens as we sample genes with ever-lower levels of codon usage bias. By sampling genes with ever-higher levels of codon usage bias, we additionally examine whether the mutagenic effect of late replication is sufficient to impact substitution rates in even the most highly expressed genes.

Materials and Methods

Orthology

Orthology information, defined by fuzzy reciprocal BLAST and Synpipe (see Clark et al. 2007), was obtained from FlyBase.

Sequence Divergence

Whether the *D. melanogaster* genome contains a class of sequence that is not under selective constraint remains contentious. There is evidence for both purifying and positive selection acting on noncoding sequences in *D. melanogaster*. Estimates of the proportion of constrained noncoding sites in *D. melanogaster* range from 22% (Bergman and Kreitman 2001) to 70% (Andolfatto 2005). It has been estimated that the genome of *D. melanogaster* contains three times as much functional noncoding sequence as protein-coding sequence and most deleterious mutations in *D. melanogaster* are thought to occur in noncoding sequence (Halligan and Keightley 2006). Indeed, certain classes of noncoding sequence appears to evolve more slowly than synonymous sites in coding sequence (Andolfatto 2005), suggesting that synonymous exonic sites might be suitable to determine whether unconstrained sites show regional variation in divergence, especially in genes with low codon usage bias.

In addition, positions 8–30 of introns $\leq$65 bp in length have been designated unconstrained based on their high levels of divergence (Halligan and Keightley 2006), while bearing the obvious disadvantage that there are relatively few of them compared with synonymous sites. Although a comparison of several classes of putatively neutral sites did not reveal a significant difference in divergence between 4-fold synonymous and intronic rates, synonymous sites were estimated to contain 7.6% more constrained positions than base pairs 8–30 of introns $\leq$65 bp in length (Parsch et al. 2010).

One caveat of the above analysis is, however, the use of all sequence within exons. Selection on exonic splice enhancers (sequences enriched near exon–intron junctions involved in the recognition of exon ends), is known to pose an additional selective constraint on coding sequence, and failure to remove these sites can lead to the true rate of divergence being underestimated, particularly in genes with a high proportion of sequence near intron–exon boundaries (see Parmley et al. 2007; Warnecke and Hurst 2007; Warnecke et al. 2008). Given the above, we consider two classes of site: positions 8–30 in small introns and synonymous sites in exon cores (i.e., away from exon–intron junctions—see below for definition). In addition, we consider the effects of selection on codon usage by considering what happens as we enrich the sample for lowly expressed genes.

Synonymous Exon Core Divergence

To calculate divergence at synonymous sites, 6 of the 12 available *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*) were considered, as divergence beyond the melanogaster group is too great for dS to be estimated accurately (see Supplementary information, Supplementary Material online in Larracuente et al. 2008). Exon nucleotide sequences for all six species were obtained from FlyBase release 5.33. Exons were matched by filtering out genes that did not have at least one isoform with identical exon numbers across all six species, and exons with corresponding positions within a gene that differed by more than 10% in length or that were absent in one or more species were discarded.

We trimmed 69 bp from each exon boundary, as these regions are subject to constraint due to selection on splice enhancers, which can lead to divergence being underestimated (see Parmley et al. 2007; Warnecke and Hurst 2007; Warnecke et al. 2008). The nucleotide sequences were then translated to amino acids, aligned in MUSCLE (Edgar 2004) and translated back to nucleotides. Sequences with internal stop codons due to missing bases in the nucleotide sequence file were removed. All exons from the same gene were then concatenated to reduce noise resulting from the use of short sequences to calculate divergence. PAML was used to calculate synonymous divergence rates using an unrooted version of the best-supported tree from Larracuente et al. (2008) (see fig. 1). As our replication timing estimates come from *D. melanogaster* Kc cells, divergence between *D. melanogaster* and its reconstructed ancestral node was considered in our analysis. We employed three models. First, one where the codon model was F3x4 and omega was permitted to vary across branches. Second, one using the same codon matrix as above but restricting to one omega across branches. Finally, one where codons were assumed equally abundant (1/61 each) and omega permitted to vary across branches. In all cases, we assume no clock and do not permit variation in omega between sites within a gene. Results for all

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

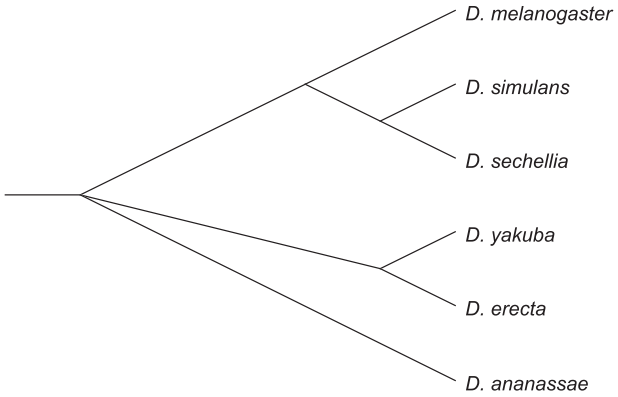

Fig. 1. Phylogenetic tree used to calculate divergence for the *melanogaster* subgroup (see Larracuente et al. 2008). Our dS values consider evolution from the common ancestor of *Drosophila melanogaster* with *D. simulans* and *D. sechellia* down the *D. melanogaster* line.

measures are all but identical. We report in the main text those for F3x4 and omega variable. Relevant results for the other measures are reported in the supplementary Material online.

Intronic Divergence

The problem of employing divergence of bases 8–30 of introns with ≤65 bp length is the limited number of introns meeting the length criterion. Rather than aligning sequences from all six species, three-way alignments (dmel/sim/sec and mel/yak/ere) were made to maximize sample size. Alignments were performed with MAFFT using the—genafpair—maxiterate 1000 settings. *Drosophila ananassae* was not considered as the alignments did not appear reliable when checked by eye. Whole introns were aligned and then trimmed before further processing. Intronic divergence was then calculated using the method of Tamura and Kumar (2002), which accounts for compositional heterogeneity. In addition, we calculated net divergence, subtracting nucleotide diversity (see below) from the estimated divergence (see Li 1977) to control for ancestral polymorphism, so as not to overestimate divergence between recently diverged species (Charlesworth 2010).

Codon Usage Bias

Codon usage bias was calculated using codonW (http://codonw.sourceforge.net). This provides a score for each gene that reflects its tendency to employ the optimal codons. Codon adaptation index (CAI) reference values (i.e., defining the optimal codons) were taken from Carbone et al. (2003).

Replication Timing

Replication times in *D. melanogaster* were determined by Schwaiger et al. (2009). Of the cell types for which replication timing data was obtained, embryonic-derived Kc cells were considered to be the closest to the germ line. These data were obtained from the NCBI Gene Expression Omnibus (GEO) in Series GSE13328, data set GSE336362, file GSM336362_Kc_replication_timing.txt. Array probe start positions for the original data set were individually converted from UCSC assembly dm2 to UCSC assembly dm3 using the UCSC liftOver tool and associated chain file dm2ToDm3.over.chain. Array probe start positions were then further converted from UCSC assembly dm3 to FlyBase release 5 using the FlyBase coordinate converter located at http://flybase.org/static_pages/downloads/COORD.html (Tweedie et al. 2009). The lifted probe positions were first purged to remove any whose position could not be converted to FlyBase release 5. Next, all cases whereby a genomic location on FlyBase release 5 had been assigned to more than one probe were identified. Replication times at each position were then compared. Where replication times did not differ between duplicate probe positions, only a single probe was retained in the final data set. Where replication times differed, replication times for each pair were plotted and their orthogonal residual from $x = y$ were calculated. From the distribution of these residuals (supplementary figs. 1 and 2, Supplementary Material online), an upper limit of 0.069 was imposed. Both duplicate probes with orthogonal residuals >0.069 were purged from the final data set. Where orthogonal residuals were ≤0.069, a mean of the two replication times was taken and assigned to a single probe at that position. Genic positions in R5.33 were extracted from file dmel-all-CDS-r5.33.fasta, available as a precomputed file from FlyBase at ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.33_FB2011_01/fasta. Replication times were assigned to genes based on identification of all probes located within a given gene. Where more than one replication time was identified for a given gene, these were tested for normality of distribution, 23% of which were found to be skewed. As such, the median of all replication times that applied to a given gene was taken. Plots of the replication time and dS as they vary across the genome are in supplementary figs. 3–7 (Supplementary Material online).

Recombination Rates

Per-gene recombination estimates based on Release 4.3 of the *D. melanogaster* genome were obtained from Larracuente et al. (2008). These rates were calculated by two different methods: regression polynomial (RP) is based on the slope of the third-order regression polynomial at the midpoint of each gene (see Hey and Kliman 2002), whereas adjusted coefficient of exchange (ACE) is calculated from the relationship between genetic and physical map positions across polytene bands, that is, on a local scale. For plots of recombination rate against chromosomal position, see supplementary figs. 8–12 (Supplementary Material online). We used the genes as reference points when comparing recombination rate against local diversity.

Polymorphism

Estimates of Tajima's Pi from parallel sequencing of Portuguese strains of *D. melanogaster* calculated in 10 kbp windows given the parameters in http://code.google.com/p/popoolation/wiki/PoPOOLationWalkthrough were obtained from Kofler et al. (2011) (see Futschik and Schlötterer 2010) and mapped to Flybase identifiers. For plots of diversity against

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

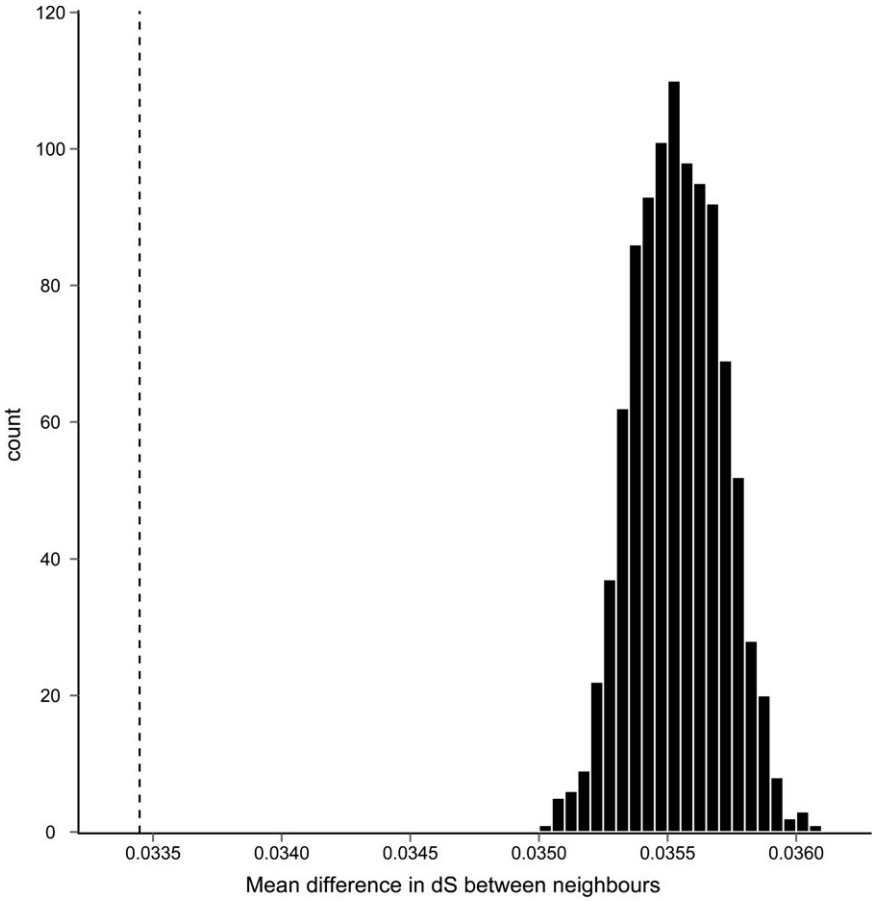

FIG. 2. The distribution of the mean difference in dS between "neighbors" in randomizations in which a gene can adopt any position along the chromosome on which it is found. Shown as the dashed line is the mean observed difference.

chromosomal position, see supplementary figs. 13–17 (Supplementary Material online).

Results

Regional Similarities in dS Are Greater Than Expected by Chance

If the regional homogeneity in mutation rates that has been suggested to exist in mammals (see e.g., Wolfe et al. 1989; Matassi et al. 1999; Smith et al. 2002; Lercher et al. 2004) is also present in *D. melanogaster*, we expect to detect local similarities in divergence at neutral sites. An autocorrelation for local diversity and divergence rates has been observed in *D. simulans* (Begun et al. 2007), although it is not clear whether this observation can be extrapolated to neutral sites in *D. melanogaster*. Given the correlation between synonymous and nonsynonymous sites (Begun et al. 2007), we might, however, expect to detect a local similarity in dS in *D. melanogaster*. This is indeed what we find.

To minimize noise in estimation, we considered genes with at least 500 bp of exonic sequence ($n = 2599$). For these long genes, we calculated the modular difference in exon core dS between the neighbors and calculated the mean of these differences. We then compared this value against the ones for randomized genomes, in which a gene could take any position on the chromosome arm that it was originally found on. For these randomized genomes, we again determined the mean difference in dS between neighbors. We find that values of adjacent genes are significantly more similar than expected based on 1000 randomizations (none of the simulations was as similar as the real comparisons: fig. 2). The same is observed when we exclude the X-linked genes ($n = 2388$) or when we exclude genes in the telomeric 3MB. Thus, we confirm that, as with mammals, neighboring genes have similar rates of evolution at synonymous sites.

Late-Replicating Sequence Has Elevated Rates of Evolution

Given the local similarities in substitution rates at weakly constrained sites, we then ask whether local rates might, as in mammals, be impacted by replication timing. If so, we would expect exon core dS values for late-replicating genes to be higher than for early-replicating genes. As above, we removed genes with less than 500 bp concatenated exon sequence length. In this set, replication timing negatively correlates with exon core dS values (Spearman's rho, hereafter referred to as rho $= -0.0564$, $P = 0.004$, $n = 2560$), that is, late-replicating sequence has increased dS (see fig. 3). Note that negative replication timing values indicate late replication. When we consider the difference between the earliest and latest replicating sequences, we find a difference of 10.81% (determined by substituting the earliest and latest replication times into

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

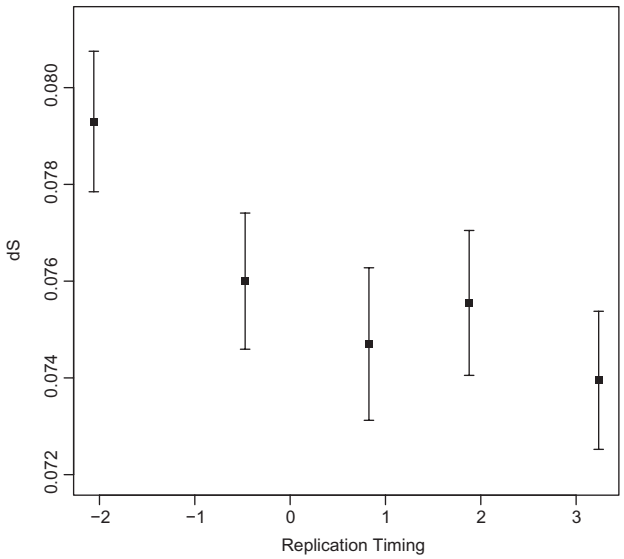

Fig. 3. Late-replicating sequences have elevated exon core dS. Plot represents median dS values for 5 bins of equal size binned by replication timing with standard error. In a quadratic fit to the raw data, the linear term ($P = 0.01$) is significant but the quadratic term is not ($P = 0.35$). The possible upturn at early-replicating domains is thus not of statistical relevance.

the equation for the linear regression between dS and replication timing). The magnitude of this effect is in accord with prior reports from mammals (Pink and Hurst 2010).

When comparing closely related species, ancestral polymorphism can lead to the true extent of divergence being overestimated (see Charlesworth 2010). To control for this, polymorphism (Pi) for the 10-kb region overlapping each gene was subtracted from dS to give a corrected net divergence value (see Li 1977). Our results are not altered when this polymorphism-corrected measure of synonymous divergence (dS − Pi) is employed (rho = −0.0505, $P = 0.0116$, $n = 2603$; difference between early and late replicating = 10.99%). Note, however, that this is not a perfect control as the diversity measure does not distinguish between classes of site. However, a partial correlation of dS against replication timing controlling for diversity also remains highly significant ($P = 0.007$). Both tests suggest that the replication time dS correlation is not owing to standing variation but reflects substitutional differences.

Analysis of introns provides qualitatively similar results to those seen for synonymous sites but not always significant. As only 23 bp of sequence from introns ≤65 bp could be used for a given intron, sequences were partitioned into 100 bins of equal size according to their replication timing. Sequences from each bin were then concatenated to give an average divergence estimate for each bin. When we consider the more distant comparator species *D. yakuba* and *D. erecta*, we find, as with synonymous sites, a significant negative relationship (rho = −0.2385, $P = 0.0169$ for *D. yakuba*; rho = −0.2195, $P = 0.0282$ for *D. erecta*) (see fig. 4). The difference in divergence between the earliest and latest replicating bins is 8.07% for *D. yakuba* and 8.27% for *D. erecta*, supporting the notion that late-replicating introns are more diverged. However, for intronic divergence between *D. melanogaster* and *D. simulans* and between *D. melanogaster* and *D. sechellia*, we observe negative but nonsignificant correlation with replication timing (rho = −0.1098, $P = 0.2767$ for *D. simulans*; rho = −0.1366, $P = 0.1754$ for *D. sechellia*). That there is considerable scatter in figure 4 may reflect, among other things, differences in replication timing between species, error in estimation of replication timing and variable levels of constraint even within the 23 bp of sequence in short introns.

Analysis of Low Codon Usage Bias Genes Suggest a 30% Difference in Mutation Rate between Early- and Late-Replicating Sequence

The above analyses suggest that late replicating, possibly neutral sequence has a higher mutation rate than early-replicating sequence. However, it is likely that the intronic sequence may yet be under selective constraint and many of the genes in our sample are likely to be affected by selection on codon usage. In

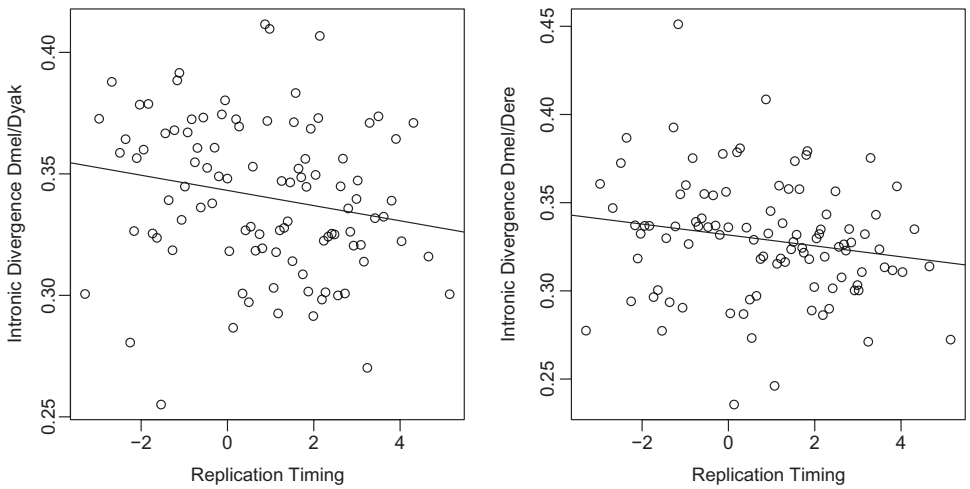

Fig. 4. Late-replicating introns have higher divergence between *Drosophila melanogaster* and *D. erecta* and between *D. melanogaster* and *D. yakuba* than early-replicating introns. Introns were split into 100 bins of equal size by replication timing and concatenated prior to calculating divergence values. This result is robust to changes in the number of bins. Mean *P*-values for one-tailed Spearman's correlation are 0.028 for *D. yakuba* and 0.038 for *D. erecta* for 75, 90, 100, 110, and 125 bins.

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

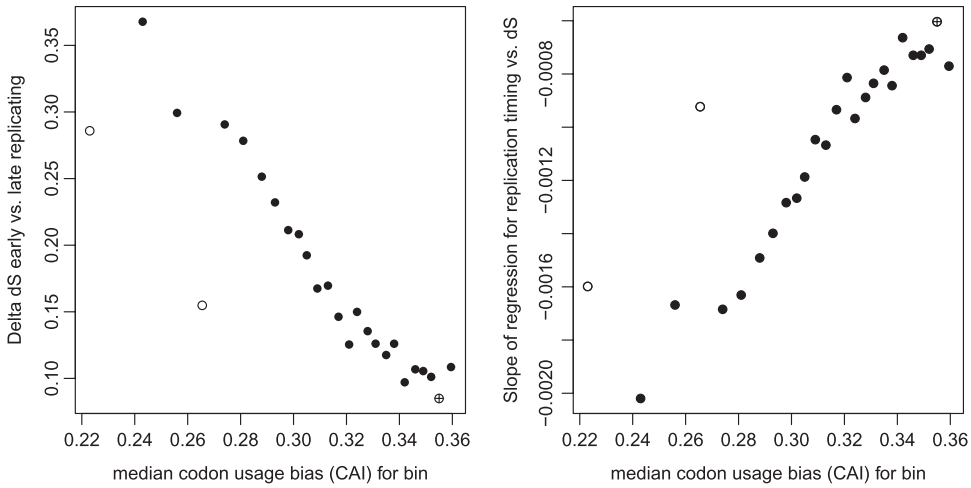

Fig. 5. High-CAI sequences mask the size of the effect of replication timing on exon core dS. Sequences are binned by exon core CAI and one bin at a time is removed, starting with the most highly biased. (*A*) indicates the percentage difference in dS between the earliest and latest replicating sequences when substituted into the equation describing the linear regression between replication timing and dS. (*B*) shows the slope of the regression line. Solid circles indicate that $P \leq 0.05$ for the regression slope. Crosshatched circles indicate $0.05 < P <= 0.1$. Empty circles indicate that $P > 0.1$.

principle, if we wish to estimate the magnitude of the difference in mutation rate between early- and late-replicating sequence, we should consider what happens when we restrict analysis to genes with low codon usage bias. Note that CAI is more appropriate than expression to control for evolutionary constraint imposed by high levels of expression, as CAI reflects all of life history and optimal codons are known to differ between life stages (see Hense et al. 2010).

Fortunately, CAI (and Fop) does not correlate with replication timing ($P = 0.9532$ for CAI; $P = 0.4731$ for Fop), indicating that there is no tendency for more highly codon-biased genes to replicate early or late. Enriching the sample for lowly expressed genes should not thus greatly skew replication timing values, possibly distorting statistics. This lack of correlation is perhaps not unexpected, as replication timing is more closely correlated with histone 4 lysine 16 acetylation (H4k16ac) than with transcription, with 30% of early-replicating genes being inactive (Schwaiger et al. 2009).

Rather than arbitrarily deciding that a certain CAI score should be considered as a cutoff for low bias, instead we consider what happens to the difference between the earliest and latest replicating sequence as we gradually remove the genes with the highest CAI scores from the analysis and consider what happens at the limit. That is, we assigned genes to 26 bins of equal size according to CAI and removed one bin at a time, beginning with the most highly biased. We then calculate the slope on the line relating replication time to dS and the percentage difference between the earliest and latest in terms of their replication timing by reference to the regression line and to two fixed time points (the maxima and the minima from the unreduced data set).

We observe that the percentage difference in dS between the earliest and latest replicating sequence increases markedly as highly codon-biased genes are removed from the set (see fig. 5*a*). Concomitantly, an increase in magnitude is also observed for the slope of the regression line, which becomes steeper the more highly biased bins are removed (see fig. 5*b*). This conclusion is robust to removal of 3 Mb from the telomeric end each chromosome (supplementary fig 18. Supplementary Material online). Employing an estimation of dS where omega is held constant across lineages also makes no difference (supplementary fig. 19, Supplementary Material online). Similarly, using a different codon matrix model makes no difference (supplementary fig. 20, Supplementary Material online). Considering the limit as CAI is ever lower the percentage difference appears to asymptote to about ~30%, while in the unfiltered sample, it is just 10.81%. We conclude that selection on codon usage partially obscures the difference between early- and late-replicating sequence.

We can also consider the opposite trend, that is, what happens as we sequentially remove the least biased (lowest CAI) genes (fig. 6). Is selection on the most highly expressed genes strong enough to counteract any increased mutation rate in late-replicating sequence? We find that the difference between early- and late-replicating sequences (or rather the slope of the regression line) is no longer significant, suggesting that selection on codon usage in the most highly expressed genes is indeed strong enough to mask underlying mutation rate differences (see fig. 6). After removing the first 2 of 13 bins, there appears to be a decline in the difference in exon core dS between the earliest and latest replicating sequences with a borderline significant regression slope (delta = 0.0841, $P = 0.0846$ after removing the first bin; delta = 0.0845, $P = 0.0986$ after removing the second bin; see fig.6), but the difference quickly diminishes thereafter. This contrasts with the previous analysis where most bins showed a significant slope (see fig. 5).

Late-Replicating Sequences Have Elevated Diversity Not Explained by Recombination

Given the observed increase in divergence in late-replicating regions, we might expect to see a similar effect on nucleotide diversity if mutation rates do indeed increase in late S-phase.

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

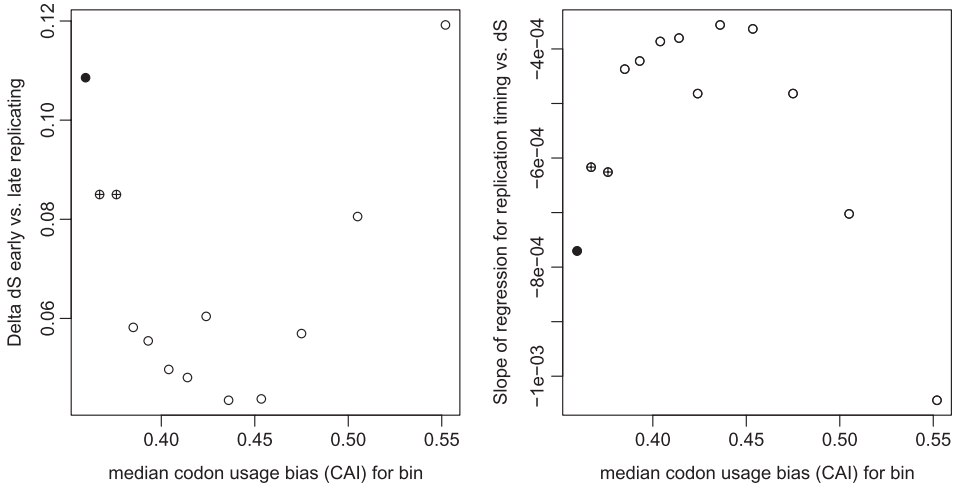

Fig. 6. Selection on codon usage may be compromised in late-replicating regions, but the effect of replicating late in S-phase on dS is only significant in the least biased subset of genes. Sequences are binned by exon core CAI and one bin at a time is removed, starting with the least biased. (*A*) indicates the difference in dS between the earliest and latest replicating sequences when substituted into the equation describing the linear regression between replication timing and dS. (*B*) shows the slope of the regression line. Solid circles indicate that $P \leq 0.05$ for the regression slope. Crosshatched circles indicate $0.05 < P <= 0.1$. Empty circles indicate that $P > 0.1$.

Estimates of nucleotide diversity depend on the study population (see Sackton et al. 2009). In the Portuguese sample, we consider median diversity (measured by Tajima's Pi, see Kofler et al. 2011) for all regions is 0.0055 compared with previous reports of 0.0063 for African flies and 0.0049 for non-African samples (see Aquadro et al. 2001). As with exon core dS and intronic divergence, nucleotide diversity is elevated in late-replicating sequences (rho $= -0.0384, P = 3.66 \times 10^{-10}$), with a 10.61% difference between late- and early-replicating regions (see fig. 7). It is interesting to note the similarity in the magnitude of the effect on diversity and synonymous divergence (prior to CAI filtering). This increase in diversity in late-replicating regions accords with findings from mammals (Stamatoyannopoulos et al. 2009; Chen et al. 2010).

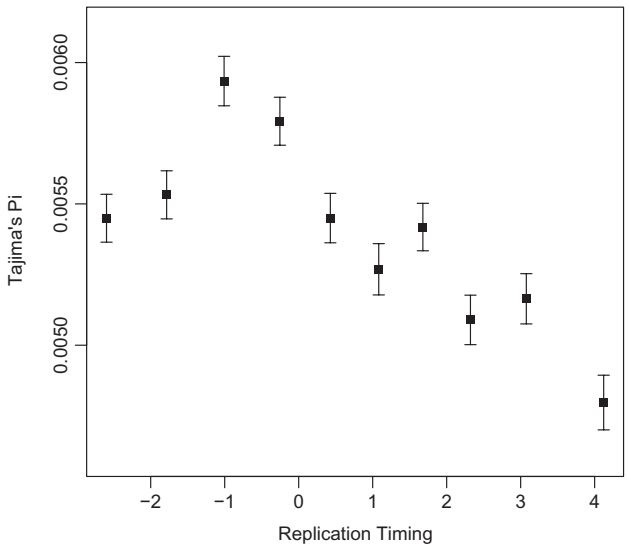

Fig. 7. Nucleotide diversity (measured by Tajima's Pi) is increased in late-replicating domains. Plot represents median Pi values for ten bins of equal size binned by replication timing with standard error.

As expected were there regional homogeneity but pan-genomic heterogeneity in the mutation rate, synonymous divergence and diversity are positively correlated (for genes with at least 500 bp of exon core sequence rho $= 0.1666$, $P < 2.2 \times 10^{-16}, n = 2603$). This accords with the prior finding that divergence between *D. melanogaster* and *D. simulans* is correlated with diversity within *D. melanogaster* (Sackton et al. 2009).

Although consistent with the possibility that late replication causes an increase in the mutation rate and thus higher levels of diversity, the above need not necessarily be indicative of a mutational effect. Recombination is also associated with increased levels of nucleotide diversity in multiple species including *D. melanogaster* (see e.g., Begun and Aquadro 1992; Lercher and Hurst 2002; Hellmann et al. 2005; Begun et al. 2007; Sackton et al. 2009; Stevison and Noor 2010). Two possible explanations for this that are not mutually exclusive have been put forward: The effect has been attributed to the greater independence between physically linked alleles in highly recombining regions (Kulathinal et al. 2008; Cutter and Choi 2010; Cutter and Moses 2011) or, alternatively, a coupling between recombination and mutation (Lercher and Hurst 2002; Hellmann et al. 2003; Kulathinal et al. 2008). As regards the former, both background selection and hitchhiking with positively selected variants cause markers in linkage disequilibrium to lose heterozygosity thus reducing diversity. In domains of high recombination the genomic span over which linkage disequilibrium can affect the fate of a given allele is much reduced, hence, diversity is expected to be higher. The possibility that, in Drosophila, recombination might itself be mutagenic (Magni and von Borstel 1962) is lacking strong support. Some analyses find no significant relationship between putatively neutral divergence and recombination (see Betancourt and Presgraves 2002; Stevison and Noor 2010) and any positive correlation between putatively neutral divergence and

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

recombination is of uncertain interpretation (Ometto et al. 2006; Kulathinal et al. 2008). Whatever the cause, we expect that domains of low recombination to be domains of low diversity. In accord with previous reports, we detect this predicted positive correlation between both ACE and RP and Pi (All chromosomes: rho = 0.3660, $P < 2.2 \times 10^{-16}$, $n = 11727$ for ACE; rho = 0.4342, $P < 2.2 \times 10^{-16}$ for RP; Autosomes only: rho = 0.5215, $P < 2.2 \times 10^{-16}$, $n = 9895$ for ACE; rho = 0.5782, $P < 2.2 \times 10^{-16}$ for RP).

If late-replicating regions tend also to be highly recombining, then the tendency for late replication to be associated with high diversity is not in itself adequate to show that late replication is mutagenic. We thus ask whether covariation with recombination rates might account for the increase in diversity we observe in late-replicating regions.

If late-replicating *D. melanogaster* genes have high rates of recombination, we might expect this to be apparent in our data set, as meiotic cross over occurs in the female germ line only, and our replication timing estimates come from Kc cells (Schwaiger et al. 2009), which are inferred to be female based on dsx splicing (see http://flybase.org/reports/FBtc0000998.html). However, presently available recombination maps have limited resolution (see supplementary figs. 8–12, Supplementary Material online), and the sign of the relationship between recombination and replication timing depends on the measure employed. RP shows a borderline significant negative association with replication timing (rho = −0.0168, $P = 0.07$, $n = 11727$) indicating increased recombination rates in late-replicating sequences, whereas ACE is significantly positively correlated with replication timing (rho = 0.0255, $P = 0.00577$) indicating decreased recombination in late-replicating sequences. The result for the former measure is consistent with the possibility of recombination being responsible for the increase in diversity in late-replicating regions, whereas the latter is not. As opposed to RP, ACE estimates recombination on a local scale (see supplementary figs 8–12, Supplementary Material online), making it the more appropriate measure here, given the relatively small span of regions with similar replication timings in *D. melanogaster* (see Schwaiger et al. 2009). Additionally, when only autosomal genes are considered, the correlation between RP and replication timing becomes nonsignificant (rho = −0.0069, $P = 0.4943$, $n = 9895$), whereas the correlation between ACE and replicating timing becomes stronger (rho = 0.0490, $P = 1.057 \times 10^{-6}$).

We therefore ask whether the correlation between diversity and replication timing is independent of recombination. Employing partial Spearman's correlation, we find that controlling for ACE in fact increases the strength of the relationship between Pi and replication timing (rho = −0.0354, $P = 0.0001$; beta = −0.0481, $P < 10^{-3}$, $n = 11727$). This increase in magnitude becomes yet more profound when only autosomal diversity is considered (rho = −0.0557, $P = 2.87 \times 10^{-8}$; beta = −0.0954, $P < 10^{-3}$, $n = 9895$). On the other hand, the effect of replication timing on the relationship between recombination and diversity is negligible (rho = 0.5215, $P < 2.2 \times 10^{-16}$; beta = 0.5257, $P < 10^{-3}$ for autosomes only; note rho and beta are all but identical), as expected given the strong correlation between recombination and polymorphism. We can therefore infer that, far from explaining the relationship between replication timing and diversity, the effects of recombination may mask part of the impact that replication timing has on nucleotide diversity. This is presumably owing to early-replicating regions that are also highly recombining.

Discussion

Contrary to previous reports based on limited data (Begun et al. 2007) and in accord with findings from mammals (Stamatoyannopoulos et al. 2009; Chen et al. 2010; Pink and Hurst 2010), we observe an increase in nucleotide diversity and putatively neutral substitution rates in late-replicating sequences in *D. melanogaster*, which is consistent with an increase in the mutation rate during late S-phase. The magnitude of this effect, between circa 10% and 30% difference, is roughly in accord with estimates from mammals (10–20%). Selection on codon usage in highly expressed genes can be enough to counteract these mutational differences.

That the same trend is seen in mammals, yeast and flies (Stamatoyannopoulos et al. 2009; Chen et al. 2010; Pink and Hurst 2010; Lang and Murray 2011) suggests that the trend for late-replicating sequence to have high mutability is a phylogenetically widespread feature in eukaryotes. Indeed, we are tempted to speculate that it may be a universal feature of eukaryotes. Evidence for a similar effect in Archea (Flynn et al. 2010) suggests that it may be even more widespread.

If the above is indeed the case we might expect a mechanism that is itself phylogenetically conserved to explain the trends. It has, for example, been suggested that elevated CpG mutation rates in mammals are accounted for by the increase in methylation observed in late-replicating genes (Chen et al. 2010). An association with methylation, however, is unlikely to provide a general model being absent from yeast and flies. Indeed the increase in CpG mutation rates through the cell cycle are matched by those at non-CpG sites (Chen et al. 2010), indicating that methylation alone is not an adequate explanation (see Mugal and Ellegren 2011).

One recently emerged hypothesis has the possibility of being generally applicable. During replication, after the replication fork has passed over an area, it is possible that one strand of what should be double-stranded DNA is just single stranded (see Discussion in Lopes et al. 2006; Chen et al. 2010). Such single-stranded DNA can be repaired by one of two mechanisms. The first uses the complementary sequence from the other product of replication. This is not especially error prone. An alternative is highly error prone translesion synthesis. In yeast, the key enzyme for this mode of synthesis, Rev1, is seen only very late in replication, possibly as a last attempt to repair single-stranded DNA before termination of cell division. This tempts the question as to whether the same trend is seen in flies. *Drosophila* has an ortholog of Rev1. Analysis of expression profile through the cell cycle would be informative.

Our analysis raises a further series of issues. Why, for example, do we report the common result (late replication

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

associated with high divergence and diversity), while Begun et al. found the opposite from a more limited analysis of chromosome 2L using 10-kb binned data? One possibility is that their replication timing data disagrees with ours. This appears not, however, to be the explanation. Identifying genes in the domains they consider to be early replicating, we find that they are also mostly early replicating in our sample (median replication time 2.254 vs. a genomic median of 0.82). Another possibility is using all sites in 10 kb windows as opposed to synonymous sites might mislead. Unfortunately, their analysis was limited in scope so there are only 16 genes longer than 500 bp for which we could obtain evidence. From this, we cannot make any definitive conclusions. However, we see no difference in the mean dS between these genes and the set of several thousand genes in our sample (mean in our sample = 0.081, mean in theirs = 0.082, $P = 0.814$). This does not lend support to the view of a higher mutation rate in early-replicating domains. Another possibility is that 2L is not representative of the genome. For this, we see some evidence. 2L is unusually early replicating (median replicating time 1.35 compared with a genomic mean of 0.82, $P < 0.00001$) and unlike the two arms of chromosome 3, 2L does not show a significant trend for dS to be reduced in early-replicating domains. This may reflect the fact that it has few truly late-replicating domains. Analysis of 2L alone is thus unlikely to provide a reliable guide to genomic trends.

Our results also have relevance for population genetical inference. It is commonly supposed that the level of diversity can be employed to understand the role of selection. Classically, low diversity is employed to identify domains subject to selective sweeps, for example, via hitchhike mapping (Harr et al. 2002). While these methods allow for diversity to vary as a function of the local recombination rate, they typically do not allow for replication timing differences. It is thus possible that the method will be prone to calling false positives for genes in domains of early replication and false negatives for genes in domains of late replication. A very early to replicate domain, for example, is expected to have low diversity regardless of any sweep. The extent of and quantitative impact of this further necessary covariate in analyses we leave to future studies. Replication timing will, we suggest, be an important parameter in future population genetical analyses.

Supplementary Material

Supplementary figures 1– 20 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Robert Kofler for providing the nucleotide diversity data, Ichiro Hiratani for help with the replication timing data, and Amanda Larracuente for providing the recombination rate estimates. L.D.H. is a Royal Society Wolfson Research Merit Award Holder, C.J.P is sponsored by MRC, and C.C.W. by The University of Bath.

References

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437:1149–1152.

Aquadro CF, DuMont VB, Reed FA. 2001. Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev.* 11: 627–634.

Begun DJ, Aquadro CF. 1992. Levels of naturally-occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* 356:519–520.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.* 5:2534–2559.

Belyakin SN, Babenko VN, Maksimov DA, Shloma VV, Kvon EZ, Belyaeva ES, Zhimulev IF. 2010. Gene density profile reveals the marking of late replicated domains in the Drosophila melanogaster genome. *Chromosoma* 119:589–600.

Belyakin SN, Christophides GK, Alekseyenko AA, Kriventseva EV, Belyaeva ES, Nanayev RA, Makunin IV, Kafatos FC, Zhimulev IF. 2005. Genomic analysis of Drosophila chromosome underreplication reveals a link between replication control and transcriptional territories. *Proc Natl Acad Sci U S A.* 102:8269–8274.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11:1335–1345.

Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in Drosophila. *Proc Natl Acad Sci U S A.* 99: 13616–13620.

Carbone A, Zinovyev A, Kepes F. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19:2005–2015.

Charlesworth D. 2010. Don't forget the ancestral polymorphisms. *Heredity* 105:509–510.

Chen CL, Rappailles A, Duquenne L, et al. (11 co-authors). 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* 20:447–457.

Clark AG, Eisen MB, Smith DR, et al. (414 co-authors). 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. *Genome Res.* 20:1103–1111.

Cutter AD, Moses AM. 2011. Polymorphism, divergence, and the role of recombination in Saccharomyces cerevisiae genome evolution. *Mol Biol Evol.* 28:1745–1754.

Donaldson AD. 2005. Shaping time: chromatin structure and the DNA replication programme. *Trends Genet.* 21:444–449.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. 2010. Evolutionary rates and gene dispensability associate with replication timing in the archaeon Sulfolobus islandicus. *Genome Biol Evol.* 2:859–869.

Futschik A, Schlötterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186:207–218.

Gilbert DM. 2002. Replication timing and transcriptional control: beyond cause and effect. *Curr Opin Cell Biol.* 14:377–383.

Grasser F, Neusser M, Fiegler H, Thormeyer T, Cremer M, Carter NP, Cremer T, Muller S. 2008. Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. *J Cell Sci.* 121:1876–1886.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16:875–884.

Harr B, Kauer M, Schlotterer C. 2002. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024

mutations in Drosophilamelanogaster. *Proc Natl Acad Sci U S A.* 99:12949–12954.
Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet.* 72: 1527–1535.
Hellmann I, Prufer K, Ji HK, Zody MC, Paabo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* 15:1222–1231.
Hense W, Anderson N, Hutter S, Stephan W, Parsch J, Carlini DB. 2010. Experimentally increased codon bias in the Drosophila Adh gene leads to an increase in larval, but not adult, alcohol dehydrogenase activity. *Genetics* 184:547–555.
Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of drosophila. *Genetics* 160:595–608.
Hiratani I, Ryba T, Itoh M, et al. (12 co-authors). 2010. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res.* 20:155–169.
Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang C-W, Lyou Y, Townes TM, Schübeler D, Gilbert DM. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* 6:e245.
Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One.* 6:e15925.
Kulathinal RJ, Bennettt SM, Fitzpatrick CL, Noor MAF. 2008. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. *Proc Natl Acad Sci U S A.* 105:10051–10056.
Lang GI, Murray AW. 2011. Mutation rates across budding yeast Chromosome VI are correlated with replication timing. *Genome Biol Evol.* 3:799–811.
Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet.* 24:114–123.
Lercher MJ, Chamary J-V, Hurst LD. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res.* 14:1002–1013.
Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* 18:337–340.
Li WH. 1977. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* 85: 331–337.
Lopes M, Foiani M, Sogo JM. 2006. Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. *Mol Cell.* 21:15–27.
MacAlpine DM, Rodriguez HK, Bell SP. 2004. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev.* 18:3094–3105.
Magni GE, von Borstel R. 1962. Different rates of spontaneous mutation during mitosis and meiosis in yeast. *Genetics* 47: 1097–1108.
Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol.* 9:786–791.
Mugal CF, Ellegren H. 2011. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol.* 12:R58.
Ometto L, De Lorenzo D, Stephan W. 2006. Contrasting patterns of sequence divergence and base composition between Drosophila introns and intergenic regions. *Biol Lett.* 2:604–607.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:343–353.
Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Mol Biol Evol.* 27:1226–1234.
Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B. 2006. Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet.* 38:1005–1014.
Pink CJ, Hurst LD. 2010. Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol Biol Evol.* 27:1077–1086.
Ranz JM, Diaz-Castillo C, Petersen R. 2011. Conserved gene order at the nuclear periphery in Drosophila. *Mol Biol Evol.* 29: 13–16.
Sackton TB, Kulathinal RJ, Bergman CM, Quinlan AR, Dopman EB, Carneiro M, Marth GT, Hartl DL, Clark AG. 2009. Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biol Evol.* 1:449–465.
Schmegner C, Berger A, Vogel W, Hameister H, Assum G. 2005. An isochore transition zone in the NF1 gene region is a conserved landmark of chromosome structure and function. *Genomics* 86:439–445.
Schmegner C, Hameister H, Vogel W, Assum G. 2007. Isochores and replication time zones: a perfect match. *Cytogenet Genome Res.* 116:167–172.
Schwaiger M, Stadler MB, Bell O, Kohler H, Oakeley EJ, Schubeler D. 2009. Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. *Genes Dev.* 23:589–601.
Smith NGC, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* 12:1350–1356.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet.* 41:393–395.
Stevison LS, Noor MA. 2010. Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis. *J Mol Evol.* 71:332–345.
Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol.* 19:1727–1736.
Tweedie S, Ashburner M, Falls K, et al. (11 co-authors). 2009. FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Res.* 37:D555–D559.
Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. *Mol Biol Evol.* 24:2755–2762.
Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9:R29.
Williams RR, Azuara V, Perry P, et al. (11 co-authors). 2006. Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *J. Cell Sci.* 119:132–140.
Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285.
Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919.

Downloaded from https://academic.oup.com/mbe/article/29/2/873/1032772 by guest on 24 April 2024