Timing of Replication Is a Determinant of Neutral Substitution Rates but Does Not Explain Slow Y Chromosome Evolution in Rodents

Catherine J. Pink and Laurence D. Hurst*

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

***Corresponding author:** E-mail: bssldh@bath.ac.uk.

Associate editor: James McInerney

Abstract

Mutation rates, assayed as substitution rates of putatively neutral sites, are highly variable around mammalian genomes: There is heterogeneity between genes, between autosomes, and between X, Y, and autosomes. The differences between X, Y, and autosomes are typically assumed to reflect the greater number of cell divisions in the male germ-line. Such an effect can neither account for within-autosome differences nor does it predict the differences between X, Y, and autosome observed in rodents. It has recently been proposed that in primates, the time during S-phase when a gene is replicated is an important determinant of neutral rates of evolution. Here we ask 1) whether we can replicate this result in rodents, 2) whether different autosomes replicate on average at different times, and 3) whether this might explain differences in their substitution rates. Finally we ask 4) whether X, Y, and autosome replicate at different times and 5) whether any difference might explain why the number of replication events alone cannot explain their substitution rates. We find that, as in primates, autosomal intronic rates of evolution increase significantly during S-phase. Different autosomes do have different average replication times, and together with rearrangement, this is a significant predictor of between-autosome differences in substitution rate. Although we find that autosomal, X-, and Y-linked genes replicate at different times, it is paradoxical that the Y-linked genes replicate latest, and replicate more often, but are not especially fast evolving. These results support the hypothesis that replication timing is an important source of substitution rate heterogeneity.

Key words: replication timing, mutation, rodents, male-driven evolution, Y chromosome evolution.

Introduction

Mutations rates, assayed as the substitution rate at putatively neutral sites, are known to vary at different scales across mammalian genomes, but the reasons for this are not well resolved. On the same autosome, genes differ in their synonymous substitution rate (Wolfe et al. 1989) with genes of similar substitution rate clustering (Matassi et al. 1999; Lercher et al. 2001), an effect that is not explained by clustering of genes with similar expression profile (highly/broadly expressed genes tending to have lower synonymous rates; Lercher et al. 2004). Domains of similarity in substitution rate appear to be defined by synteny blocks, genes within a block having more homogeneity than between blocks (Malcom et al. 2003; Webster et al. 2004).

At a more gross level, we see striking differences between chromosomes. Not only are there differences between X, Y, and autosome (Shimmin et al. 1993; Chang et al. 1994; Smith and Hurst 1999; Makova and Li 2002; Sandstedt and Tucker 2005; Goetting-Minesky and Makova 2006; Bachtrog 2008; Pink et al. 2009) but there are also differences between autosomes (Lercher et al. 2001; Ebersberger et al. 2002; Malcom et al. 2003; Gaffney and Keightley 2005; Pink et al. 2009). In part, the explanation for the differences between X, Y, and the average autosomal rate is thought to reflect different numbers of cell divisions owing to different times spent in male versus female germ-lines (e.g., see Crow 1997a, 1997b; Hurst and Ellegren 1998; Li, Yi, and Makova 2002; Ellegren 2007). This theory, the theory of male-driven evolution (Miyata et al. 1987), assumes both that the majority of mutations arise as errors during DNA replication and that, per replication, these errors are uniformly distributed throughout the genome. Mutational variability should therefore reflect only differences in the number of replications sequences undergo. Given that in longer lived species, maintenance of spermatogonia increases the number of germ-line cell divisions in males relative to females, the Y chromosome, which is restricted to males, might be expected to have a higher substitution rate than the autosomes, which are only exposed to additional male germ-line replications half of the time. In turn, autosomes should evolve faster than the X chromosome that spends only one-third of its time in the male germ-line. Recent evidence from rodents has, however, shown that the number of replication events is unable to explain observed differences between chromosomal classes, both in exonic synonymous substitution rates and intronic rates (Pink et al. 2009). For both classes of sequence, estimates of the extent of the male bias (α), based on a model presuming that the number of replications is the sole determinant of neutral substitution rates, varied significantly depending on which two chromosomal classes were considered (X vs. autosomes, X vs. Y, Y vs. autosome). Indeed, in strict contradiction of the hypothesis, the autosomes were found to have

a similar, if not higher substitution rate than the Y chromosome.

As previously reported (Matassi et al. 1999; Lercher et al. 2001; Malcom et al. 2003), Pink et al. (2009) also found considerable between-autosome variability in putatively neutral substitution rates. Neither this observation nor the discrepant estimates of α are consistent with between-chromosomal variability in mutation rates being predominantly determined by the number of germ-line cell divisions. While Pink et al. (2009) proposed a recombination-associated substitution effect as a source of the higher autosomal rate of evolution than that of the nonrecombining Y chromosome, the source of the between-autosomal variability in substitution rates remains unresolved. Although a recombination-associated substitution effect can explain some part of the between-autosomal gene variation ($r^2 = 0.035$, $P = 5 \times 10^{-5}$ [Pink et al. 2009]), it fails to explain any of the between-autosome variation (Pink CJ, Hurst LD, unpublished data).

Given that the number of DNA replications cannot account for variability in substitution rates between the chromosomal classes, what else might have an affect? Recent evidence from primates suggests that later replicating regions of the genome have higher rates of neutral divergence and nucleotide diversity than regions replicating earlier (Stamatoyannopoulos et al. 2009). Stamatoyannopoulos et al. (2009) postulate that the effect may be owing to a slowing or stalling of replication late in S-phase, possibly owing to a depletion of the deoxynucleotide triphosphate (dNTP) pool or difficulty negotiating heterochromatized templates. The slower speed of replication would in turn mean that DNA would be unwound, in single stranded format, for longer, leaving it more prone to mutation. However, the mechanism is by no means well resolved. Speed of fork progression appears to be a dynamic feature of replication related to other factors such as dNTP availability (Malínsky et al. 2001; Anglana et al. 2003) and origin density (Conti et al. 2007; Courbet et al. 2008), and it is not yet fully understood how these vary temporally across S-phase. Regardless of the mechanistic uncertainties, importantly, replication timing tends to be a relatively fixed property of a genomic domain, remaining stable from cell cycle to cell cycle (Jackson and Pombo 1998), with GC-rich, gene-rich regions tending to replicate earlier than AT-rich, gene-poor, or heterochromatic regions (Woodfine et al. 2004; Karnani et al. 2007; Hiratani et al. 2008).

Replication timing data are now available at a 5.8-kb probe density across all three chromosomal classes in mouse (Hiratani et al. 2008), the only mammalian species to our knowledge with such data available for the Y chromosome. We therefore ask whether timing of replication is also related to substitution rates in rodents and further, whether it can account for the previously observed inter-autosomal variability. We finally ask whether any differences in replication timing between the three chromosomal classes (X, Y, and autosome) are causative of differences in substitution rate, previously thought attributable to differences in the number of replications in the two germ-lines, and whether controls for replication timing may resolve the previous discrepancies in models used to estimate the extent of the male mutation bias.

Methods

Calculation of Intronic Substitution Rates

Sequence extraction, orthology definition, and filters were as previously described in Pink et al. (2009). In brief, autosomal and X-linked intronic sequences were obtained from University of California–Santa Cruz (UCSC) genome browser using the July 2007 assembly for *Mus musculus* and the November 2004 assembly for *Rattus norvegicus*. Orthologous autosomal and X-linked genes were identified initially from a set of Mouse Genome Informatics (MGI)-defined mouse–rat orthologs (Eppig et al. 2007) and further strictly defined as true orthologs based on similarity of exon number, exon phase, and chromosomal class (autosomal, X, or Y). Rat Y-linked intronic sequences were obtained using the methods previously described (for accession numbers, see Pink et al. 2009 supplement) and subjected to a BlastN search against the mouse genome to identify orthologous mouse Y-linked introns, for which intronic sequences were downloaded from UCSC (Karolchik et al. 2004).

Orthologous intronic sequences were aligned individually using LAGAN (Brudno et al. 2003). By reference to a set of hand-aligned mouse–rat introns (Chamary and Hurst 2004), alignments of a length greater than 1.16 times the length of the longest sequence or that contained more than 0.084 indels per aligned base were purged from the analysis, these numbers representing the best aligned 95% of the data. 30 bp were then removed from the ends of each alignment to control for conservation of splice sites and indels were removed from the alignments. First introns were eliminated from the analysis, these known to be unusually slow evolving (Keightley and Gaffney 2003; Chamary and Hurst 2004).

Two data sets were then produced. The first unfiltered data set comprised all alignments that passed the above filters. The second data set was further purged of all introns thought to be evolving under purifying selection, possibly owing to the inclusion of unannotated exons within intronic sequence. This involved subjecting alignments in the first data set to a test for clusters of conserved bases, potentially indicative of hidden functional sites, in which any introns with a lower number of switches (between conserved and nonconserved residues or vice versa) per base than predicted by a linear model were eliminated from the data set (see Pink et al. 2009 for details). For both data sets, introns of the same gene were then concatenated and the rate of intronic divergence (K_i) was estimated and corrected for multiple hits according to the model of Tamura and Kumar (2002).

Assignment of Chromosomal Locations

Positions of genes on the mouse genome were defined by the terminal 5′ and 3′ bp of the coding sequence. These positions were obtained from annotations of the July 2007 assembly (mm9). As mouse replication timing data were

assigned genomic coordinates based on the February 2006 assembly (mm8), the stand-alone liftOver utility and associated chain file mm9ToMm8.over.chain, both obtained from UCSC, were used to convert positions between builds.

Replication Timing Data

Replication timing data for mouse cell lines prior to differentiation were downloaded from http://www.replicationdomain.org (Hiratani et al. 2008). Positive values were indicative of early replication and negative values were indicative of replication later during S-phase. Four data sets were available. Three comprised replication times derived from embryonic stem cells (ESCs), whereas the fourth set of replication times was derived from induced pluripotent stem cells (iPS). Although the three ESC lines can be regarded as replicate data sets, the same is not necessarily true of the iPS data. Therefore, to justify the inclusion of data derived from iPS cells, for each chromosome a Spearman's correlation was performed on the raw data for each possible pairwise comparison between the four data sets, enabling a comparison of the strength of correlations within the ESC data to those between any of the ESC data and the iPS data. Correlations in chromosomal replication timing between pairwise ESC lines were no stronger than correlations between any of the ESC lines and the iPS line (supplementary fig. 1, Supplementary Material online), confirming the finding by Hiratani et al. (2008) that replication profiles of iPS cells were indistinguishable from other ESCs. We therefore treated the four cell lines as replicates.

Assignment of Genic and Chromosomal Replication Times

For each orthologous gene, all replication times obtained from the four cell lines that applied to any part of the gene were identified based on an overlap of the positions of the probe used to calculate replication times and the limits of the coding sequence. A mean of these replication times was then assigned to the gene. From this data set of orthologous genes with both substitution rate and replication time data available, the median intronic substitution rate and median replication time across all genes located on each chromosome were used for analysis at the chromosomal level; 95% confidence intervals were determined from 1,000 bootstraps.

Controls for Germ-Line Expression

To our knowledge, no rodent germ-line expression data that are not limited to the advanced stages of gametogenesis are currently available. However, strand asymmetry in the rates of some substitution types has resulted in an excess of G and T over C and A on the coding strand in mammals (Green et al. 2003; Mugal et al. 2009). This asymmetry is higher in transcribed than in flanking intergenic sequence (Green et al. 2003; Mugal et al. 2009), and transitions between equal and skewed base composition are clearly associated with the start and end points of transcription (Touchon et al. 2003, 2004; Polak and Arndt 2008). Together, these observations are strongly suggestive of a germ line transcription-associated source. Further, the extent of this skew has been found to correlate with expression level in ubiquitously expressed genes (Majewski 2003). As such genes are more likely to be expressed in the germ-line than tissue-specific genes, we therefore used the extent of G + T skew as a proxy for germ-line expression rate. For each mouse intron, the numbers of A, T, C, and G were determined, and the extent of G + T was skew calculated as the ratio $[(G + T) - (A + C)]/(G + C + T + A)$ (Majewski 2003).

Rearrangement Index

Using the final samples of orthologous genes for either the filtered or unfiltered data set as appropriate, for a given mouse autosome, two genes located on that autosome were randomly selected. Whether the rat orthologs of this pair of mouse genes were located on the same rat autosome or on two different rat autosomes was then established. For the focal mouse autosome, this process of randomly sampling pairs of mouse genes and identifying the location of their rat orthologs was repeated 10,000 times and the number of occasions on which the rat orthologs were located on two different autosomes was counted (n). Division of this count by the number of random samplings (i.e., n/10,000) generated an index of between-autosomal rearrangement for the focal mouse-autosome. This method was then applied to each mouse autosome to generate its rearrangement index, such that chromosomes having undergone extensive between-autosomal rearrangements were assigned high rearrangement indices, whereas low rearrangement indices were assigned to autosomes that have remained relatively collinear since their common ancestor with rat. Full details of the sample sizes, counts, and rearrangement indices are supplied in supplementary table 1 (Supplementary Material online). Due to the criteria by which orthologs were selected, both genes having to be located on a chromosome of the same class (X, Y, or autosome), it was not possible to apply this method of quantifying between-chromosomal rearrangements to the sex chromosomes as all orthologs would, by definition, be located on the same chromosome, giving rise to an index of 0. We note that this rearrangement index does not quantify the extent of intrachromosomal rearrangements such as inversions.

Results

We generated two data sets: one subject to a filter for introns thought to contain clusters of sites under selective constraints and a second data set not subject to this filter. As our findings do not, for the most part, qualitatively differ between the two data sets, for brevity, we present here the results from the more conservative, filtered data set. This comprised 4,378 autosomal genes (18.7 Mb), 133 X-linked genes (622 Kb), and 3 Y-linked genes (5.5 kb). Results from the unfiltered data set can be found in the supplement.

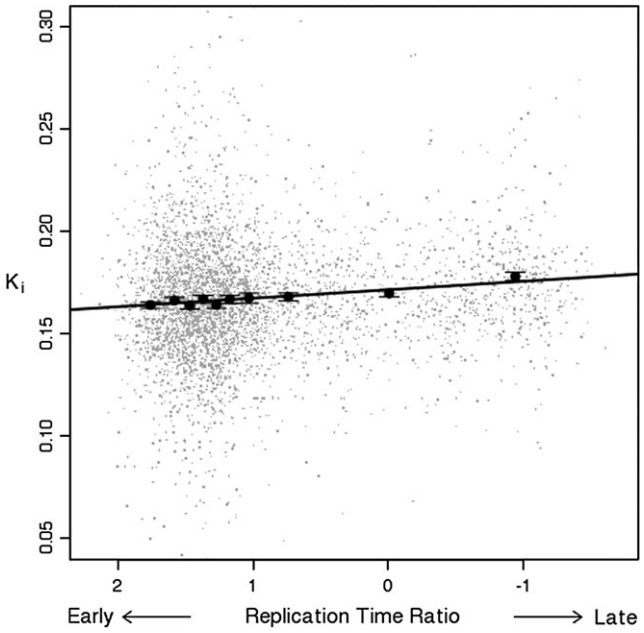

Fig. 1. Intronic substitution rate increases with later timing of replication across autosomal genes. Spearman's $\rho = -0.0901$, $P = 2.3 \times 10^{-9}$. Also shown are bin averages (±1 standard error of the mean) for equally sized bins. Regression line is for all data, not bin means. Note that the *y* axis scale results in some outlying data points lying outside the plot area.

Replication Time Correlates with Intronic Rates of Evolution

We first asked whether, at the genic level, timing of replication is related to putatively neutral substitution rates. Confirming the previously reported trend, seen in primates (Stamatoyannopoulos et al. 2009), in rodents, there is a significant relationship between timing of replication and intronic substitution rates across both autosomal genes (Spearman's $\rho = -0.0901, P = 2.3 \times 10^{-9}$; fig. 1) and X-linked genes (Spearman's $\rho = -0.2188, P = 0.0114$). Note that due to the structure of the data, late-replicating sequences are assigned negative timing values, so an increase in any variable during S-phase yields a negative correlation. For figures, data have been plotted on a reversed *x* axis so as to visually show this increase over time. Using the regression ($K_i = -0.00440$ (replication time) + 0.1717) to predict K_i from the replication times of the first and last genes to replicate, we see an expected 10.5% increase in rates of evolution during S-phase. However, this is considerably lower than the 22% increase in divergence reported across primate temporal replication states (Stamatoyannopoulos et al. 2009).

GC Content Does Not Explain Why Early Replicating Genes Evolve Slowly

Parenthetically, it is interesting to note that the replication time effect runs opposite to a nucleotide-level mutability effect. Consistent with previous work (Woodfine et al. 2004; Hiratani et al. 2008), we observe a significant, strong correlation between GC content and replication timing across autosomal genes, such that GC-rich sequences replicate early, whereas sequences that are GC-poor replicate late (Spearman's $\rho = 0.3153$, $P < 2.2 \times 10^{-16}$). GC-rich sequences should, thus, evolve slowly owing to replication timing effects. Indeed, we find a significant, albeit weak, negative relationship (Spearman's $\rho = -0.0525$, $P = 0.00051$; see also Pink et al. 2009), although it should be noted that this relationship is sensitive to the data set used (see Supplementary Results, Supplementary Material online). By contrast, synonymous substitution rates have been found to covary positively with GC content (Hurst and Williams 2000), and further, CpG dinucleotides are known to be mutable especially when methylated (Coulondre et al. 1978). Whether a replication time effect will interfere with attempts to infer patterns of germ-line methylation indirectly via examination of CpGs (e.g., Meunier et al. 2005; Rollins et al. 2006; Sigurdsson et al. 2009) is unknown.

Given that we find that GC-rich sequences replicate early and are somewhat slow evolving, we therefore asked whether this might account for the relationship between replication timing and intronic substitution rates. We find that when we account for GC content, the strength of the relationship between replication time and intronic substitution rate is somewhat weaker but remains significant (partial Spearman's $\rho = -0.0777$, $P = 0.0010$), suggesting that this effect is not modulated by GC content. Conversely, most of the relationship between GC and intronic substitution rates is explained by GC-rich sequences being early replicating ($\rho^2 = 0.003$ for uncontrolled analysis, $P = 0.0005$; $\rho^2 = 0.0006$ for the partial correlation controlling for replication time, $P = 0.042$).

Expression Rate Does Not Explain Lower Rates of Evolution of Early Replicating Genes

In mammals, early replication has been associated with gene expression (Holmquist 1987; Woodfine et al. 2004). It might therefore be the case that the lower substitution rate we observe in earlier replicating genes could be explained by features relating to gene expression. Consistent with this hypothesis, we find a significant correlation between replication time and germ-line expression rate, as assayed by nucleotide skew (Spearman's $\rho = 0.0969$, $P = 1.3 \times 10^{-10}$) with highly expressed genes replicating earlier. However, whether we might, a priori, expect such genes to have lower rates of evolution is unclear, not least because previous evidence has been conflicting. At synonymous sites, the strength of the relationship has ranged from weakly negative (Lercher et al. 2004) to non-existent (Duret and Mouchiroud 2000) and although a significant correlation between intronic rates of evolution and several measures of expression rate has previously been reported in humans (Webster et al. 2004), we are unable to replicate this with our rodent data (Spearman's $\rho = 0.0209$, $P = 0.166$). It is therefore unsurprising that using a partial correlation, a significant correlation between replication time and substitution rate remains when controlling for germline expression rate (partial Spearman's $\rho = -0.0926$, $P = 0.0010$). We conclude that the lower substitution rate of earlier replicating genes is not attributable to higher levels of gene expression.

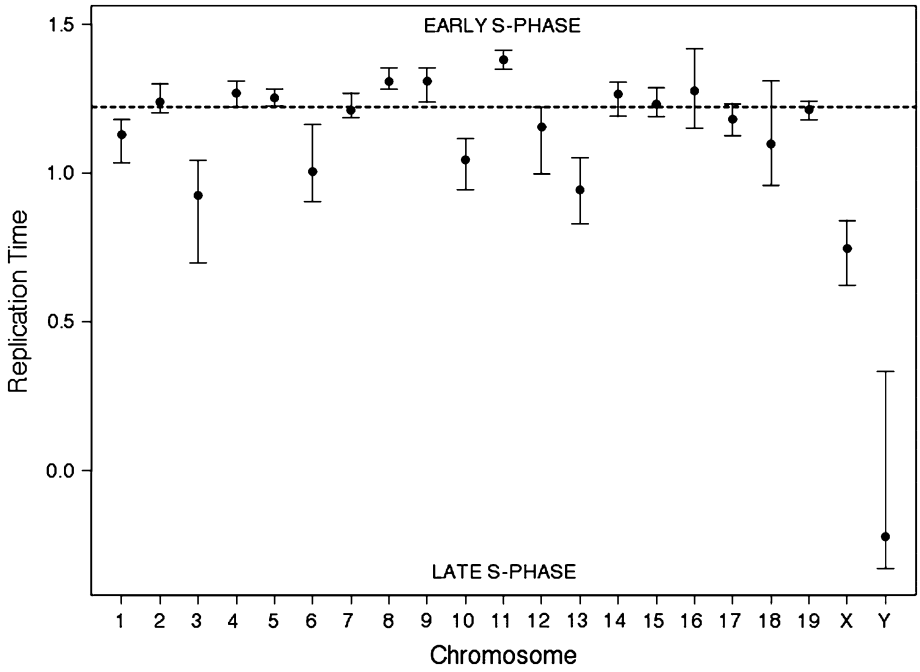

FIG. 2. Median chromosomal replication time (±95% confidence intervals) for each of the 19 mouse autosomes and the two sex chromosomes. Horizontal line represents the median across all autosomal genes. There are, on average, significant differences in the timing of replication time of different autosomes (Kruskal–Wallis, $P < 2.2 \times 10^{-16}$) and between the three chromosomal classes; X, Y, and autosome (Kruskal–Wallis, $P < 2.2 \times 10^{-16}$).

Differential Timing of Replication, In Part, Explains Inter-autosomal Variation in Substitution Rates

The theory of male-driven evolution (Miyata et al. 1987) suggests that, if the majority of mutations arise as errors during DNA replication, then at the chromosomal level, variation in substitution rates is predominantly determined by differences in the number of replications in each germ line, more occurring in males than in females of longer lived species. This theory would therefore predict that the autosomes should all evolve at the same rate as, on average, they pass through each germ-line with equal frequencies. However, significant differences in rates of autosomal evolution (Malcom et al. 2003; Gaffney and Keightley 2005; Pink et al. 2009) suggest that the number of replications is not the sole determinant of autosomal mutation rates.

Given that we find an increase in genic rates of evolution as replication progresses through S-phase, we therefore asked whether this effect extends to the inter-chromosomal level. First, we asked whether, on average, the autosomes replicate at different times. We find that there is considerable heterogeneity between autosomes in their replication timing (Kruskal–Wallis, $P < 2.2 \times 10^{-16}$; fig. 2).

We then asked whether these differences in replication time between autosomes were related to differences in autosomal substitution rates. There is a correlation between replication timing of autosomes and their intronic substitution rate, but whether this is significant depends somewhat on exactly how the data are handled ($\rho_{max} = -0.547$, $\rho_{min} = -0.216$, $P_{min} = 0.017$, $P_{max} = 0.373$). Overall, we observe about a 4.5% difference in mean rates between the earliest and latest replicating autosomes. However, we have previously found that, for reasons unknown, highly rearranged mouse autosomes have high substitution rates compared with those that have not undergone substantial inter-chromosomal rearrangements (Pink et al. 2009). Given the strength of this relationship ($r^2 = 0.6063$, $P = 0.0001$; fig. 3*a*), it must be considered alongside any other parameter under investigation as a cause of between-autosomal variation in substitution rates, in this instance, timing of replication.

We therefore first ask whether replication timing and amount of interchromosomal rearrangement are independent parameters. As we find no correlation between the two variables (Spearman's $\rho = 0.0288$, $P = 0.907$) and in a generalized linear model, in which both are predictors of autosomal substitution rates, there is no significant interaction term ($P = 0.3495$), we conclude that this is indeed the case. Excluding an interaction from the generalized linear model, we then find that together, rearrangement and replication timing can explain a striking 70% of between-autosomal variation in substitution rates ($r^2 = 0.732$, $P = 2.689 \times 10^{-5}$). Although both parameters contribute significantly to this relationship in the filtered data, rearrangement appears to be the dominant predictor ($P = 1.94 \times 10^{-5}$ for rearrangement compared with $P = 0.0147$ for replication timing). This can be seen also by considering how well replication time predicts the residuals of the plot of rearrangement index against autosomal intronic rates (fig. 3*b*). Although the significance of replication timing as a co-predictor of autosomal substitution rates in the unfiltered data is sensitive to how autosomal averages were

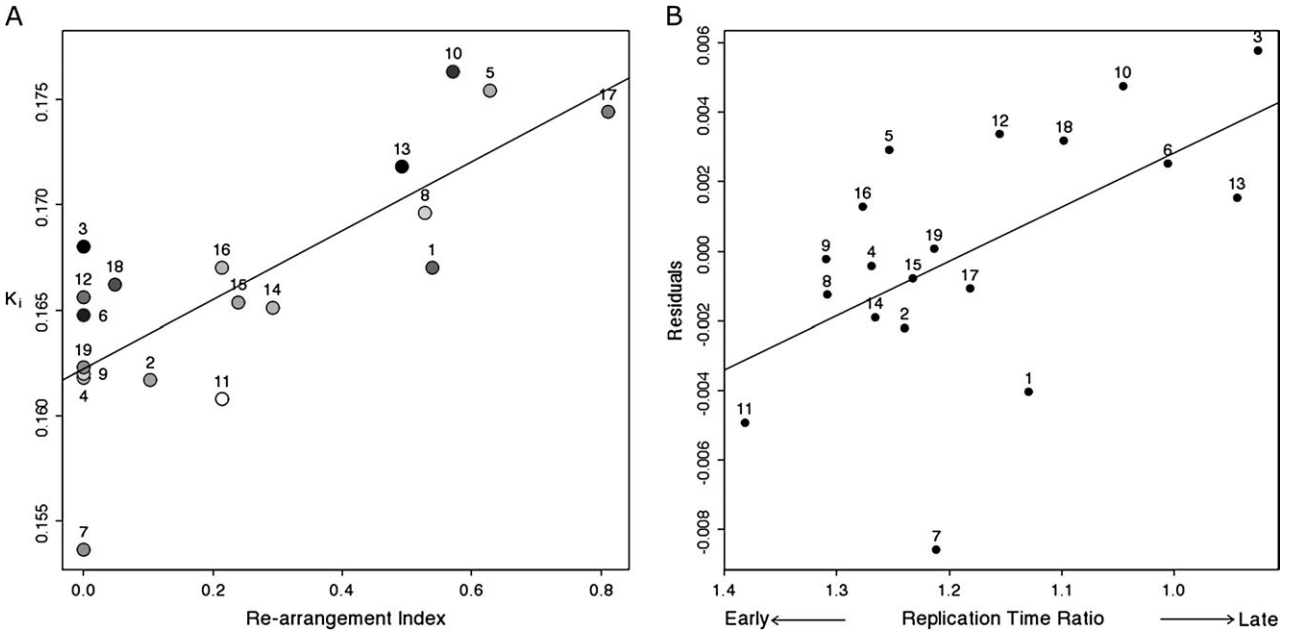

FIG. 3. Rearrangement and replication time predict autosomal intronic rates. The intronic substitution rates of the 19 (labeled) mouse autosomes (a) are significantly predicted by the amount of rearrangement the autosome has undergone ($r^2 = 0.606$, $P = 0.0001$, regression line shown) and timing of replication (residuals test $r^2 = 0.318$, $P = 0.0119$, darker points being indicative of later replication timings). Note the tendency for later replicating autosomes to sit above the line and early replicating ones to sit below. This is further illustrated in (b), a plot of the residuals for (a) against replication time. Together, these two parameters can explain over 70% of between-autosomal variation in substitution rates (generalized linear model $r^2 = 0.732$, $P = 2.69 \times 10^{-5}$).

calculated ($P = 0.0299$ for means, $P = 0.1865$ for medians), given that we do observe significant relationships in most of our analyses suggests that replication timing should be considered as a covariate in future analyses investigating the causes of variation in autosomal rates of evolution.

Mean Replication Time of X, Y, and Autosomal Genes Are Different but Controlling for Replication Time Does Not Account for Discrepancies in Estimates of α

We recently found that, contrary to the predictions of the theory of male-driven evolution, Y-linked introns have a rate of evolution that is, at most, on a par with those of the autosomes, if not somewhat lower (Pink et al. 2009). This is also true considering synonymous sites (McVean and Hurst 1997; Smith and Hurst 1999; Pink et al. 2009). More generally, estimates of α, the degree of male bias, derived using the method of Miyata et al. (1987), are not mutually compatible when using data from the three possible pairwise comparisons (X and autosomes, Y and autosomes, and X and Y). Given that later replication timing elevates substitution rates both at the genic and the autosomal level, these discrepancies might be accounted for if the autosomes replicate later during S-phase than the sex chromosomes. Do then autosomal, X-, and Y-linked genes replicate at different times and are autosomal genes on average late replicating, compared with those on the Y?

We find that genes located on each of the three chromosomal classes do replicate, on average, at significantly different times (Kruskal–Wallis test, $P < 2.2 \times 10^{-16}$). Contrary to the above hypothesis, however, autosomal genes replicate earliest during S-phase, followed by X-linked genes, with Y-linked genes replicating later in S-phase (median replication times: autosomes = 1.224, X = 0.747, Y = −0.223; fig. 2).

It is worth noting that our small sample of Y-linked genes were derived from two BACs and, as such, are positioned close together and therefore subject to similar regional effects, including replication time. It is therefore feasible that the difference in replication time we observe across our three chromosomal class samples might have arisen from our Y-linked sample being located in a particularly late-replicating domain. However, the distribution of replication times of our sample genes relative to all probes for a given chromosome (supplementary fig. 2, Supplementary Material online) shows that this is not the case, with sample genes being clustered in earlier replicating sequence on all chromosomes.

Given the above result, it is to be expected that the addition of replication time as a covariate will not resolve the discrepant estimates of α. Y-linked genes should have a very fast rate of evolution both because they undergo more replication events and because they are relatively late replicating. To understand the quantitative impact of replication time on estimates of α, we performed a covariate-controlled analysis of a form previously reported (Pink et al. 2009).

In order to control for replication time in estimation of the extent of male bias in the mutation rate, we imposed

Table 1. Estimates of α Controlling for a Single Timing of Replication.

Class Comparison	Regression	Replication Time Predictor	Predicted K_i	Ratio		α	
				Original	Controlled for RT	Original	Controlled for RT
	$K_{Auto} = 0.1717 - 0.0044 \times RT$		$K_{Auto} = 0.172$	0.8369	0.8524	2.9160	2.5887
X to Auto	$K_x = 0.1458 - 0.0113 \times RT$	$RT_Y = -0.073$	$K_x = 0.1466$				
Y to Auto	$K_{Auto} = 0.1717 - 0.0044 \times RT$	$RT_Y = -0.073$	$K_{Auto} = 0.172$	0.9521	0.9274	0.9087	0.8646
Y to X	$K_x = 0.1458 - 0.0113 \times RT$	$RT_Y = -0.073$	$K_x = 0.1466$	1.1377	1.0879	1.2218	1.1380

NOTE.—RT is the replication time and K is the intronic substitution rate of X, Y, and autosomes.

a single replication time across all three chromosomal classes and calculated the magnitude of α using the predicted rate of evolution of each chromosomal class at this time. Because the limited sample size available for the Y chromosome prevented use of a regression of Y-linked genes, we used the mean replication time across Y-linked genes to predict both autosomal and X-linked substitution rates using the equation for the regression line of replication time as a predictor of K_i across all autosomal and X-linked genes, respectively. The ratio of K_Y to the predicted estimate of K_{Auto} was then inserted into the equation of Miyata et al. (1987) to determine α_{YA} from $(K_Y/K_{Auto})/(2 - (K_Y/K_{Auto}))$. Similarly, the ratio of K_Y to the predicted K_X was used to calculate α_{YX} from $(2(K_Y/K_X))/(3 - (K_Y/K_X))$. Finally, the predicted estimate of K_{Auto} relative to the predicted K_X was used to calculate α_{XA} from $3(K_X/K_{Auto}) - 4)/(2 - 3(K_X/K_{Auto}))$. We find that controlling for replication time fails to reconcile a to a single estimate (table 1).

Discussion

Current theory suggests that at the genome-wide level, errors introduced during DNA replication are the primary source of new mutations. An assumption of this theory is that the number of replications is the key determinant of variation in rates of evolution. However, we find that, across autosomal genes, where the number of replications is the same, the timing of replication is a significant predictor of rates of evolution. Further, we find that replication timing, in conjunction with rearrangement, is a significant predictor of autosomal rates of evolution and that together, these two parameters can explain 70% of between-autosomal variation in substitution rates.

However, we find that although the sex chromosomes replicate, on average, later than the autosomes, they do not exhibit the elevated rates of evolution that might be expected if a later timing of replication is associated with a higher input of substitutions. In fact, given that the Y chromosome undergoes an increased number of germ-line cell divisions relative to the autosomes and that Y-linked genes replicate, on average, later during S-phase, we might have expected their rate of evolution to be substantially greater than that of autosomal genes. However, this we do not find, with the Y-linked genes, if anything, evolving possibly slower than autosomal genes ($K_{Auto} = 0.1676 > K_Y = 0.1595$, although significance and magnitude is sensitive to the filters applied to the data set [see Pink et al. 2009]). It is therefore unsurprising that controlling for differences in replication time across the three chromosomal classes fails to cause the three estimates of α to converge. We previously suggested that an effect of recombination promoting neutral substitutions (either owing to direct mutational effects or owing to biased gene conversion-like processes) may be an important modulator of substitution rates (Pink et al. 2009). That the Y chromosome is nonrecombining and Y-linked genes evolve slower than expected both when considering replication time and number, only further reinforces this hypothesis.

This result aside, our results have one potential important corollary. At the genic level, replication timing appears to be an important determinant of substitution rates both in rodents (as shown here) and in primates (Stamatoyannopoulos et al. 2009). If this relationship also holds true in other species, then prior estimates of α that utilize small sample sizes are almost inevitably going to be quantitatively inaccurate. To be more precise, for estimates of α to be inaccurate, all that would be needed is that the replication timing of the sequence from one comparator chromosomal class be different from that of the other. This would be particularly acute for α derived from the X-to-autosomal comparison as this comparison is extremely sensitive to the ratio of rates of evolution. Even small inaccuracies in the measurement of substitution rates on either of these chromosomal classes, stemming from a biased sample with respect to position and consequently replication time, would therefore be amplified in inaccurate estimation of α. The problem is potentially even more profound for analyses that compare one Y-linked gene with its X-linked homolog, where, given tiny sample sizes, a major difference in replication timing of the two genes could greatly skew any estimate. If, as we find here, Y-linked sequences generally replicate later than those on the X chromosome, and if this in turns accelerates their evolutionary rate, the finding that Y-linked sequences are fast evolving relative to those on the X chromosome is to be expected, regardless of any differences in the number of replication events. A priori, therefore, our general expectation is that estimates of the magnitude of the male mutation bias derived from an X to Y comparison, employing the equation of Miyata et al. (1987), are likely to provide overestimates.

Our observations might also explain why the synteny-block appears to represent a unit of homogeneity in mutation rate variation (Malcom et al. 2003; Webster et al. 2004). Replication domains, large regions of similarly timed

replication clusters, can range in size from hundreds of kilobases (Karnani et al. 2007) to several megabases (Hiratani et al. 2008). In contrast, the scale of mutational variation has been demonstrated to be no larger than 1Mb (Gaffney and Keightley 2005). It is therefore possible that genomic rearrangements might have moved regions of similarly timed replication with associated similarity in substitution rate, into a genomic landscape differing in replication time and therefore substitution rate. An early replicating block of sequence with a slow rate of evolution, for example, could move into a domain of fast evolving sequence or vice versa. If the event was relatively recent, or if the domains brought their replication timings with them, heterogeneity between synteny blocks would result.

The analysis here comes with at least one important caveat. It is possible that the replication timing data we used in this analysis does not accurately reflect replication timings in the germ-line. The data we use, possibly superior to the somatic data of used by Stamatoyannopoulos et al. (2009), comprise replication times derived from pluripotent cells as a proxy for germ-line replication times. Given that we do observe correlations between timing of replication and other genomic features is suggestive that the data we use does reflect germ-line replication times. However, it is known that differentiation is related to temporal changes in replication for as much as 20% of the genome (Hiratani et al. 2008). Although the relationship between replication timing and gene expression is not fully understood, it has been suggested that in ESCs, lineage-specific genes may be transcriptionally silent but retain RNA polymerase promoter occupancy and as such replicate early. Upon differentiation, the transcriptional potential of these silent lineage-specific genes is removed and replication occurs later (Azuara et al. 2006; Farkash-Amar et al. 2008). During the process of gametogenesis, it is therefore likely that some regions of the genome, particularly those containing transcriptionally silent genes, would undergo such shifts in replication time. Such changes would not necessarily be conserved between oogenesis and spermatogenesis nor be distributed uniformly across the three chromosomal classes. Incorrect assignment of germ-line replication times to any of the three chromosomal classes would therefore affect relationships with substitution rates and controls for the estimation of α.

We also suppose that replication time effects have the same mutational effect on X, Y, and autosome. Might it be that the sex chromosomes are exposed to a different replication environment during S-phase? Although the formation of the XY body in the male germ-line might represent one such condition, this is unlikely to have an effect because it forms during meiotic prophase, after DNA replication has been completed, and could not therefore differentially influence the effect of replication timing on substitution rates between the chromosomal classes. Alternatively, such an effect might be female specific, involving X inactivation whereby, on average half of the time, one X chromosome is subject to transient germ-line X inactivation and subsequently replicates late during S-phase. However, this cannot account for the slow Y-linked rate of evolution, relative to that of the autosomes, because neither chromosomal class would be affected.

Assuming these caveats to be of minor importance, our results provide evidence in support of replication timing as a source of genomic variation in substitution rates and can potentially explain the previously enigmatic variation in substitution rates between synteny blocks. Although these effects only deepen the mystery of why Y-linked sequence in rodents is not especially fast evolving, more generally, it opens the possibility that all prior calculations of the extent of the male mutation bias, assuming as they do that number of replication events alone is the important determinant, are likely to be wrong. The extent to which prior estimates have misled will depend on the magnitude of the replication timing effect and the difference in timing between the sequences employed. In addition to the possible influence of recombination on differences between X, Y, and autosomes (Pink et al. 2009), in the absence of corrective data, our results provide a further reason to strongly caution against the use of Miyata's equations. Further they argue against the use of single genes or clustered genes in estimation of the impact of the number of germ-line divisions on the mutation rate in male and female germ-lines without adequate control for replication time effects.

Supplementary Material

Supplementary figures 1 and 2, table 1, and other materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

C.J.P. is funded by a Medical Research Council Capacity Building Studentship; L.D.H is a Royal Society Wolfson Research Merit Award Holder. We thank Tobias Warnecke, Adam Eyre-Walker, and an anonymous reviewer for helpful comments.

References

Anglana M, Apiou F, Bensimon A, Debatisse M. 2003. Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* 114: 385–394.

Azuara V, Perry P, Sauer S, et al. (11 co-authors). 2006. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol*. 8:532–538.

Bachtrog D. 2008. Evidence for male-driven evolution in *Drosophila*. *Mol Biol Evol*. 25:617–619.

Brudno M, Chuong D, Cooper G, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. 13:721–731.

Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol*. 21:1014–1023.

Chang BH, Shimmin LC, Shyue SK, Hewett-Emmett D, Li WH. 1994. Weak male-driven molecular evolution in rodents. *Proc Natl Acad Sci USA*. 91:827–831.

Conti C, Saccà B, Herrick J, Lalou C, Pommier Y, Bensimon A. 2007. Replication fork velocities at adjacent replication origins are

coordinately modified during DNA replication in human cells. *Mol Biol Cell*. 18:3059–3067.
Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* 274:775–780.
Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, Debatisse M. 2008. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature* 455:557–560.
Crow JF. 1997a. Molecular evolution—who is in the driver's seat? *Nat Genet*. 17:129–130.
Crow JF. 1997b. The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA*. 94:8380–8386.
Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol*. 17:68–74.
Ebersberger I, Metzler D, Schwarz C, Pääbo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet*. 70:1490–1497.
Ellegren H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc R Soc B Biol Sci*. 274:1–10.
Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. The Mouse Genome Database Group. 2007. The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res*. 35:630–637.
Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, Yakhini Z, Simon I. 2008. Global organization of replication time zones of the mouse genome. *Genome Res*. 18:1562–1570.
Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. *Genome Res*. 15:1086–1094.
Goetting-Minesky MP, Makova KD. 2006. Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *J Mol Evol*. 63:537–544.
Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*. 33:514–517.
Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schübeler D, Gilbert DM. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol*. 6:e245.
Holmquist GP. 1987. Role of replication time in the control of tissue-specific gene expression. *Am J Hum Genet*. 40:151–173.
Hurst LD, Ellegren H. 1998. Sex biases in the mutation rate. *Trends Genet*. 14:446–452.
Hurst LD, Williams EJ. 2000. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene* 261: 107–114.
Jackson DA, Pombo A. 1998. Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *J Cell Biol*. 140:1285–1295.
Karnani N, Taylor C, Malhotra A, Dutta A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res*. 17:865–876.
Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 32:493–496.
Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci USA*. 100:13402–13406.
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, Williams EJ, Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol*. 18:2032–2039.
Li WH, Yi SJ, Makova K. 2002. Male-driven evolution. *Curr Opin Genet Dev*. 12:650–656.
Majewski J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet*. 73:688–692.
Makova KD, Li WH. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416:624–626.
Malcom CM, Wyckoff GJ, Lahn BT. 2003. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol*. 20:1633–1641.
Malínsky J, Koberna K, Stanek D, Masata M, Votruba I, Raska I. 2001. The supply of exogenous deoxyribonucleotides accelerates the speed of the replication fork in early S-phase. *J Cell Sci*. 114: 747–750.
Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol*. 9:786–791.
McVean GT, Hurst LD. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* 386:388–392.
Meunier J, Khelifi A, Navratil V, Duret L. 2005. Homology-dependent methylation in primate repetitive DNA. *Proc Natl Acad Sci USA*. 102:5471–5476.
Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol*. 52: 863–867.
Mugal CF, von Grünberg H-H, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol*. 26:131–142.
Pink CJ, Swaminathan SK, Dunham I, Rogers J, Ward A, Hurst LD. 2009. Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biol Evol*. 1:13–22.
Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5′ end of human genes. *Genome Res*. 18:1216–1223.
Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH. 2006. Large-scale structure of genomic methylation patterns. *Genome Res*. 16:157–163.
Sandstedt SA, Tucker PK. 2005. Male-driven evolution in closely related species of the mouse genus Mus. *J Mol E*. 61: 138–144.
Shimmin LC, Chang BH, Li WH. 1993. Male-driven evolution of DNA sequences. *Nature* 362:745–747.
Sigurdsson MI, Smith AV, Bjornsson HT, Jonsson JJ. 2009. HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res*. 19:581–589.
Smith NG, Hurst LD. 1999. The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* 152: 661–673.
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.
Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol*. 19:1727–1736.
Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res*. 32:4969–4978.

Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett*. 555:579–582.

Webster MT, Smith NG, Lercher MJ, Ellegren H. 2004. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol Biol Evol*. 21:1820–1830.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, Carter NP. 2004. Replication timing of the human genome. *Hum Mol Genet*. 13:191–202.