J Mol Evol (2005) 61:315–324
DOI: 10.1007/s00239-004-0287-1

JOURNAL OF MOLECULAR EVOLUTION

Codon Bias and Noncoding GC Content Correlate Negatively with Recombination Rate on the Drosophila X Chromosome

Nadia D. Singh, Jerel C. Davis, Dmitri A. Petrov

Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, California 90305-5020, USA

Received: 23 September 2004 / Accepted: 10 March 2005 [*Reviewing editor:* Dr. Richard Kliman]

Abstract. The patterns and processes of molecular evolution may differ between the X chromosome and the autosomes in *Drosophila melanogaster*. This may in part be due to differences in the effective population size between the two chromosome sets and in part to the hemizygosity of the X chromosome in Drosophila males. These and other factors may lead to differences both in the gene complements of the X and the autosomes and in the properties of the genes residing on those chromosomes. Here we show that codon bias and recombination rate are correlated strongly and negatively on the X chromosome, and that this correlation cannot be explained by indirect relationships with other known determinants of codon bias. This is in dramatic contrast to the weak positive correlation found on the autosomes. We explored possible explanations for these patterns, which required a comprehensive analysis of the relationships among multiple genetic properties such as protein length and expression level. This analysis highlights conserved features of coding sequence evolution on the X and the autosomes and illuminates interesting differences between these two chromosome sets.

Key words: Drosophila — Codon bias — Recombination

Correspondence to: Nadia D. Singh; *email:* ndsing@stanford.edu

Introduction

In species with chromosomal sex determination, one sex is generally heterogametic for the X chromosome (XY or XO) while the other sex is homogametic (XX). For the sake of simplicity, we refer to the heterogametic sex as ''males'' and the homogametic sex as ''females.'' This heterogamety in males is the foundation for several differences between the X chromosome and the autosomes, which ultimately affect rates and patterns of molecular evolution. One population genetic difference between the X and the autosomes is effective population size. Assuming a 1:1 sex ratio, there are only three X chromosomes for every four autosomes, suggesting that the effective population size of the X chromosome should be lower than that of the autosomes. However, if the variance in reproductive success is greater in males than in females, this may increase the effective population size of the X chromosome. In sufficiently extreme cases, this could lead to a greater effective population size of the X chromosome relative to the autosomes (Caballero 1995; Laporte and Charlesworth 2002). A chromosome with a larger effective population size should exhibit a greater efficacy of natural selection on weakly adaptive or mildly deleterious mutations.

Another implication of chromosomal sex determination is that males are hemizygous for the X chromosome. As a consequence, X-linked alleles are immediately visible to natural selection in the heterogametic males. The exposure of X-linked alleles to selection in hemizygous individuals may lead to a more rapid accumulation of favorable mutations on the X chromosome relative to the autosomes due to

the enhanced efficacy of natural selection on recessive or partially recessive mutations (Charlesworth et al. 1987).

Further differences between the X and the autosomes stem from the evolution of the sex chromosomes from an ancient pair of autosomes (Charlesworth 1991). During this transition, the dosage of X-linked genes was reduced twofold relative to the ancestral, autosomal state, and this reduction in dosage was likely deleterious for many genes. To offset the deleterious effects of a twofold reduction in gene copy number, elaborate mechanisms of dosage compensation have evolved in several different systems (for review see Marin et al. 2000) which involve transcriptional regulation of X-linked genes. These differences in expression level and transcriptional regulation of X-linked versus autosomal genes may have important implications for the evolution of coding sequences on these chromosomes.

Heterogamety of the X chromosome in males may also help shape the gene complements of the X and the autosomes. For instance, it was recently discovered that there is a relative dearth of genes with male-biased expression on the X chromosome of Drosophila as compared to the autosomes (Parisi et al. 2003). In addition, retroposed new genes in Drosophila disproportionately originate on the X chromosome but tend to be inserted ectopically on the autosomes (Betran et al. 2002). In mammals, there also appear to be high rates of origination of retroposed genes from X-linked genes, but in contrast to the situation in Drosophila, newly retroposed genes seem to be disproportionately recruited to the X chromosome (Emerson et al. 2004).

Here we investigate whether systematic differences between the X chromosome and the autosomes affect the properties of the genes residing on these chromosomes, with a specific focus on patterns of codon bias evolution. Codon bias, or the unequal usage of synonymous codons in protein coding sequences, is generally thought to enhance translational efficiency. In genes with higher codon bias, the presence of preferred codons corresponding to the most abundant tRNAs (Shields et al. 1988) is believed to increase the efficiency and/or fidelity of translation (Akashi and Eyre-Walker 1998; Akashi et al. 1998; Bulmer 1991). It has been previously demonstrated that in Drosophila, X-linked genes have significantly higher codon bias than autosomal genes (N.D. Singh, J.C. Davis, and D.A. Petrov, unpublished data). Here we further examine differences in codon usage patterns on the X and the autosomes of the *D. melanogaster* genome. In particular, we explore the well-studied relationship between codon bias and recombination rate, analyzing the X chromosome and the autosomes separately.

Previous studies indicate that codon bias and recombination rate are positively correlated in Drosophila (Kliman and Hey 1993). This relationship was initially attributed to Hill-Robertson effects; if codon bias in Drosophila were under selection for translational efficiency, then the efficacy of this selection should increase with increased recombination, leading to increased codon bias in regions of high recombination. This is because regions of reduced recombination are generally subject to weaker purifying selection and less effective positive selection due to linkage among sites (Aquadro 1997; Begun and Aquadro 1992; Hill and Robertson 1966). However, given that 21 of 22 optimal codons in Drosophila end in G or C, the positive correlation between codon bias and recombination rate in Drosophila can also be attributed to a recombination-associated bias in patterns of neutral point substitution towards increased GC (Marais et al. 2001, 2003; Singh et al. 2004). As a result, while there is considerable debate regarding the role of Hill-Robertson interference in codon usage on the autosomes of the *D. melanogaster* genome (Hey and Kliman 2002; Kliman and Hey 1993, 2003; Marais et al. 2001, 2003), there is a consensus that both Hill-Robertson effects and neutral substitutional biases should lead to a positive correlation between codon bias and recombination.

Contrary to this prediction, we document a strong, negative correlation between codon bias and recombination rate on the X chromosome. One possible explanation for this pattern is that it is an indirect correlation, mediated by a third variable correlating with both codon bias and recombination rate. To test this hypothesis, we systematically analyzed the relationships among several genic parameters: protein length, codon bias, recombination rate, expression level, gene density, and rate of protein evolution on both the autosomes and the X chromosome.

Our data highlight the importance of treating the X and the autosomes separately in analyses of coding sequence evolution, as it is clear that there are marked differences in rates and patterns of evolution between these two chromosome sets. Our data also reveal interesting similarities and differences between the chromosomes with respect to both codon usage and relationships among other genie properties. Several relationships that we found were unexpected, which provide interesting avenues of future study.

Materials and Methods

Recombination Estimates

A list of all genes (615) localized in both the physical and genetic maps in Release 3 of the *D. melanogaster* genome was kindly provided by Flybase (David Sutherland, personal communication). A third-order polynomial curve was fitted to the genetic distance as a

function of physical distance of these genes for each chromosomal arm ($R^2 \geq 0.96$ for all arms) after visually identifying and removing outliers (n = 3, 3, 2, 3, and 0 outliers on chromosome arms 2L, 2R, 3L, 3R, and X, respectively). A thirds-order polynomial curve was fitted to the genetic distance as a function of physical distance for each chromosomal arm ($R^2 \geq 0.96$ for all arms), and recombination (cM/Mb) was calculated as the derivative of this polynomial at a given nucleotide coordinate. Recombination rates on the X chromosome were not multiplied by 4/3 to correct for differences in effective population size between the X and the autosomes.

Because the X chromosome has fewer genes than the autosomal chromosomes combined, we partitioned recombination rates into low, medium, and high categories such that we would have roughly equally-sized bins for our X chromosome data. We defined "low" recombination as greater than 0.27 cM/Mb but less than 2.93 cM/Mb, "medium" recombination as greater than or equal to 2.93 but less than 3.9 cM/Mb and "high" recombination greater than or equal to 3.9 cM/Mb. These bin definitions were used in partitioning the autosomal genes as well, but because the autosomes experience lower recombination than the X, we included an additional category for autosomal genes, less than 0.27 cM/Mb. For the same reason, there is a dearth of genes in "high" recombination category for autosomal genes, which may in part explain deviations from overall patterns in this category.

In addition, we obtained eight other estimates of recombination for 12,265 genes in the *D. melanogaster* genome; ACE, R_{TE}, HK-w, HK-p, CC99, KH93, CK00, and MMDO1 (http://pbil.univ-lyonl.fr/datasets/). The adjusted coefficient of exchange (ACE) (Kindahl 1994) estimates local recombination rates based on the relationship between the physical and genetic maps across cytological intervals and relies on DNA content estimates from Sorsa (1988) deduced from optimal densities of cytological bands. R_{TE} (Hey and Kliman 2002) uses transposable elements as markers on the cytological map. The genetic map is plotted against the physical map for each cytological chromosome section and chromosome arm; local recombination rate is estimated as the slope of this relationship measured over eight points flanking the cytological section, HK-w (Hey and Kliman 2002) also uses the slope of the relationship between the genetic map position and DNA position (from Release 2 of the *D. melanogaster* genome) across eight markers, though for this measure the markers are 493 genes with independently determined genetic map locations. These 493 genes were also used to estimate recombination (HK-p) by fitting a 4^{th} order polynomial curve to genetic map position as a function of physical position for each chromosome arm; the derivative of this polynomial is used as an estimator of local recombination rate (Hey and Kliman 2002).

Like R_{TE}, CC99 (Carvalho and Clark 1999) takes a sliding window approach to estimating local recombination rates and uses the cytological map as the basis for the physical map, though for this metric the window size was nine cytological bands as opposed to 8 cytological markers. KH93 (Kliman and Hey 1993) uses genes mapped to both polytene chromosome position and genetic map position and estimates recombination as the derivative of a 4^{th} or 5^{th} order polynomial. CK00 (Comeron and Kreitman 2000) estimates recombination as the derivative of a polynomial function relating the quantity of DNA for each cytological position and the cytological map position for each chromosome arm. Finally, MMD01 (Marais et al. 2001) uses all 892 genes that had been localized on both the physical maps (Release 1 of the *D. melanogaster* genome) and genetic maps and estimated recombination as the derivative of a 2^{nd} order polynomial curve describing genetic location as a function of physical location.

Coding Sequences and Codon Usage

We retrieved coding sequences for all genes in Release 3.2 (Flybase) of the *D. melanogaster* genome that were not located in telomeric regions (sections 1,21, 60–61 and 100) as defined by Bridges (1935). Of these 12,614 genes, 10,481 are located on the autosomes and 2133 are located on the X chromosomes. Genes mapped to heterochromatic contigs were not included in our analysis. For some genes, there were several transcripts listed; for these genes, we only included the first transcript listed in our data set, and both the protein length and codon bias estimates are based solely on this first listed transcript. For each gene, we calculated the frequency of optimal codons (FOP) based on optimal codons as defined by Duret and colleagues (Duret and Mouchiroud 1999).

K_A

For each gene in the *D. melanogaster* genome, we performed a nucleotide blast of the unannotated *D. pseudoobscura* genome (downloaded from the Baylor College of Medicine Human Genome Sequencing Center at http://www.hgsc.bcm.tmc.edu/projects/drosophila/) to identify orthologous genes. We aligned the orthologous proteins using pairwise BLAST following a previously published protocol (Conery and Lynch 2001) and retained only those genes for which greater than 60% of both protein sequences was alignable. We calculated K_A between orthologous genes using the codeml program from the PAML package (Yang 1997) letting all parameters vary.

Expression Estimates

We used expressed sequence tag (EST) counts as a rough indicator of expression level, as reported by Hey and Kliman (2002) (http://lifesci.rutgers.edu/~heylab). These data were compiled from the Drosophila Gene Index (DGI) of The Institute of Genome Research (http://ww w. tigr.org/tdb/dgi), which is a catalog of multiple EST datasets. As such, these EST counts do not reflect spatial or temporal variations in expression pattern and, accordingly, should be regarded only as crude estimates of overall expression.

Gene Density

We estimated gene density using the "Genes Per Kilobase" (GPK) metric following the example set by Hey and Kliman (2002). For each gene, we counted the number of genes included partially or completely within a 20-kb window centered on the midpoint of a given transcript. Gene coordinates for all genes were taken from header information for each gene in Release 3.2 of the *D. melanogaster* genome (Flybase).

GC Content of Noncoding Regions

For each gene in the *D. melanogaster* genome, we retrieved all intronic sequence, provided that the combined length of all introns for a particular gene exceeded 200 bp. In addition, we retrieved both 5′ and 3′ untranslated regions for all genes, provided that these regions exceeded 200 bp. We also retrieved 1000 bp "upstream" sequence for genes that were separated by more than 2 kb from their nearest neighbor 5′ of their transcription start site, as well as 1000 bp "downstream" sequence for genes that were separated by more than 2 kb from their nearest neighbor 3′ of the transcription termination site. Finally, we retrieved all "remaining" sequences that were more than 1 kb away from transcribed sites (and thereby not included in our "upstream" or "downstream" categories). GC content of each fragment in our intron, 5′ UTR, 3′ UTR, upstream, downstream, and remaining categories was estimated excluding any N's in the sequence. For our partial correlation analysis among optimal codon frequency (FOP), recombination rate, and noncoding GC content, we took the

average of the GC content of intronic, upstream, and downstream sequences for all genes that had a GC content measurement for at least one of those categories.

Results and Discussion

Codon Usage and Recombination Rate

It is well documented that codon bias is positively correlated with recombination rate in *D. melanogaster* (Comeron et al. 1999; Hey and Kliman 2002; Kliman and Hey 1993; Marais et al. 2001, 2003). This relationship was initially attributed to Hill-Robertson effects of increased efficacy of selection in areas of increased recombination (Comeron et al. 1999; Hey and Kliman 2002; Kliman and Hey 1993). Because in areas of reduced recombination, fixation probabilities of novel mutations will be affected by the fixation probabilities of mutations at linked sites, the ability of natural selection to discriminate among variants will be compromised with increased linkage among sites (Aquadro 1997; Begun and Aquadro 1992; Hill and Robertson 1966).

However, it was recently proposed that the relationship between codon bias and recombination rate on the autosomes is due to a recombination-associated bias in background substitutional patterns (Marais et al. 2001, 2003; Singh et al. 2004). Background substitutional patterns are defined as the profile of nucleotide and insertion/deletion substitutions that result from either mutation or fixation of novel mutations due to forces other than selection at the level of genetic function, including biased gene conversion or selection on a larger scale, such as at the level of regional GC content (Singh et al. 2004). Because 21 of the 22 preferred codons in Drosophila end in G or C, a recombination-associated background substitutional bias towards increased GC could generate the observed positive correlation between recombination rate and codon bias. Recent analysis of background substitutional patterns on the autosomes of the Drosophila genome does, in fact, suggest that a recombination-associated bias towards increased GC is modulating codon usage patterns on the autosomes (Singh et al. 2004), and that Hill-Robertson effects need not be invoked to explain the overall positive correlation between recombination rate and codon bias on these chromosomes.

Using optimal codon frequency (FOP) as a metric for codon bias, we confirmed that codon bias exhibits a weak but significant positive correlation with recombination rate (Kendall's $\tau = 0.039, p < 0.0001$) on the autosomes of the *D. melanogaster* genome (Fig. 1, Table 1). Though this positive correlation has been reported before (Comeron et al. 1999; Hey and Kliman 2002; Kliman and Hey 1993; Marais et al. 2001, 2003), our data indicate that it is limited to the autosomes

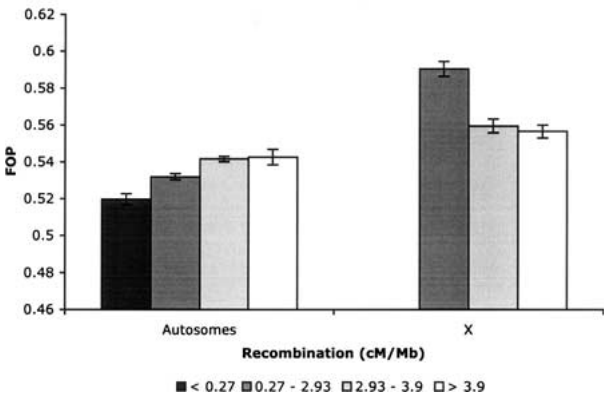

Fig. 1. Frequency of optimal codons and recombination rate, as compared between the X chromosome and the autosomes in *D. melanogaster*. As recombination rates are higher on the X chromosome, an additional recombination category for autosomes was included, corresponding to recombination below the minimum recombination estimate on the X chromosome. Error bars denote standard error.

Table 1. Kendall pairwise correlation coefficients for the autosomes and the X chromosome

Variable 1	Variable 2	Autosomes	X chromosome
FOP	Recombination	**0.039**	**−0.123**
FOP	Protein length	**−0.098**	**−0.090**
FOP	Expression level	**0.242**	**0.214**
FOP	Gene density	**0.113**	**0.239**
FOP	Rate of protein evolution	**−0.459**	**−0.408**
Recombination	Protein length	−0.007	0.006
Recombination	Expression level	−0.021	**0.090**
Recombination	Gene density	**0.056**	−0.007
Recombination	Rate of protein evolution	**−0.025**	0.065
Protein length	Expression level	**0.201**	**0.158**
Protein length	Gene density	**−0.161**	−0.128
Protein length	Rate of Protein Evolution	**−0.037**	−0.024
Expression level	Gene density	**0.077**	**0.095**
Expression level	Rate of protein evolution	**−0.221**	−0.204
Gene density	Rate of protein evolution	**0.044**	0.032

Note. Values in bold denote significance at $p < 0.05$ (Bonferroni-corrected).

(Fig. 1, Table 1). In contrast, codon bias is significantly negatively and much more strongly correlated with recombination rate on the X chromosome (Kendall's $\tau = 0.123$, $p < 0.0001$) (Fig. 1, Table 1). Notably, this correlation is not generated entirely by genes in highly recombining areas, as codon bias and recombination rate of X-linked genes are significantly negatively correlated within the lowest recombination category (0.27–2.93 cM/Mbp) (Kendall's $\tau = -0.087$, $p = 0.004$). In addition, this correlation is not an artifact of our estimates of recombination rate on the X chromosome, as all eight other estimates of recombi-

nation (see Materials and Methods) are also negatively correlated with recombination rate, six of them significantly so (Kendall's $\tau \leq -0.05, p \leq 0.03$ for ACE, HK-p, CC99, KH93, CK00, MMDO1).

It has been previously noted that the relationship between codon bias and recombination rate is non-linear, and that there appears to be a recombination threshold (of 2cM/Mbp) above which the positive relationship between these two factors no longer holds (Kliman and Hey 2003). Though our recombination categories do not delineate genomic regions precisely above and below the suggested threshold, our data do hint at the possibility that the relationship between codon bias and recombination rate is indeed non-linear (Fig. 1). For both the autosomes and the X chromosome data, the increase and decrease in optimal codon frequency, respectively, with recombination rate do appear to level off at the highest recombination category.

One possible explanation for this negative correlation between codon bias and recombination rate on the X chromosome is that it resulted from an indirect correlation with some other feature associated with the genes residing on this chromosome. To test this hypothesis, for each gene in the *D. melanogaster* genome we retrieved all available information on all additional correlates of codon bias known to date: protein length, gene density, rate of protein evolution, and expression level. Partial correlation analysis revealed that the negative correlation between recombination rate and codon bias is not mediated by these other genetic properties, as the correlation remains highly significant (Kendall partial correlation $\tau = -0.190$, Bonferroni-corrected $p = 0.0001$) (Table 2) after controlling for all additional factors. In fact, controlling for these features leads to an even stronger negative correlation between recombination rate and codon bias on the X chromosome (Tables 1, 2).

Table 2. Kendall partial correlation coefficients for the autosomes ($n = 5421$ gene) and the X chromosome ($n = 889$ genes)

Variable 1	Variable 2	Autosomes	X chromosome
FOP	Recombination	**0.049**	**−0.190**
FOP	Protein length	**−0.315**	**−0.321**
FOP	Expression level	**0.234**	**0.221**
FOP	Gene density	**0.051**	**0.197**
FOP	Rate of protein evolution	**−0.431**	**−0.394**
Recombination	Protein length	0.014	−0.093
Recombination	Expression level	**−0.043**	**0.147**
Recombination	Gene density	**0.050**	0.041
Recombination	Rate of protein evolution	−0.015	−0.022
Protein length	Expression level	**0.255**	**0.201**
Protein length	Gene density	**−0.183**	**−0.142**
Protein length	Rate of protein evolution	**−0.111**	−0.078
Expression level	Gene density	**0.108**	0.043
Expression level	Rate of protein evolution	**−0.098**	**−0.107**
Gene density	Rate of protein evolution	**0.068**	**0.134**

Note. Values in bold denote significance at $p < 0.05$ (Bonferroni-corrected).

GC Content of Coding and Noncoding Sequences

Given that 21 of 22 optimal codons in *D. melanogaster* end in G or C, the decrease in optimal codon frequency with increased recombination of the X chromosome is effectively a decrease in exonic GC content with recombination. Accordingly, the negative correlation between recombination rate and codon bias on the X chromosome could be the result of a recombination-associated bias in background substitutional patterns towards decreased G and C. To test this hypothesis, we examined the base composition of noncoding DNA in the *D. melanogaster* genome. For both the X chromosome and the autosomes, GC content of coding (as estimated through FOP) and noncoding sequences covary. GC content of noncoding DNA increases slightly with increased recombination on the autosomes (Fig. 2a) while decreasing markedly with increased recombination on the X chromosome (Fig. 2b). This is true for all types of noncoding sequences: within genes (introns and UTRs), near genes (upstream and downstream regions), as well as in regions far from genes (remaining regions).

A three-way partial correlation analysis revealed that the strength of the relationship between recombination rate and FOP in coding sequences is similar in strength to the relationship between recombination and GC content of noncoding sequences. For autosomal regions, FOP and overall GC content (see Materials and Methods) correlate to similar degrees with recombination rate, though the correlation between GC content and recombination rate is not statistically significant (Kendall partial correlations FOP vs. recombination $\tau = 0.06$, GC content vs. recombination $\tau = 0.05$ Bonferroni-corrected $p = 0.04$ and 0.12, respectively, $n = 1704$ genes). On the X chromosome, the magnitudes of the correlations are also similar (Kendall partial correlations FOP vs. recombination $\tau = -0.114$, GC content vs. recombination $\tau = -0.217$, Bonferroni-corrected $p < 0.03$, both comparisons, $n = 509$ genes). Likewise, GC content of coding and noncoding sequences are highly and significantly correlated (Kendall partial correlation $\tau = 0.278$ and 0.177, for autosomes and X chromosome, respectively, Bonferroni-corrected $p < 0.001$, both comparisons).

It is important to note, however, that the relationships between recombination rate and noncoding GC content may not be strictly linear. For autosomal

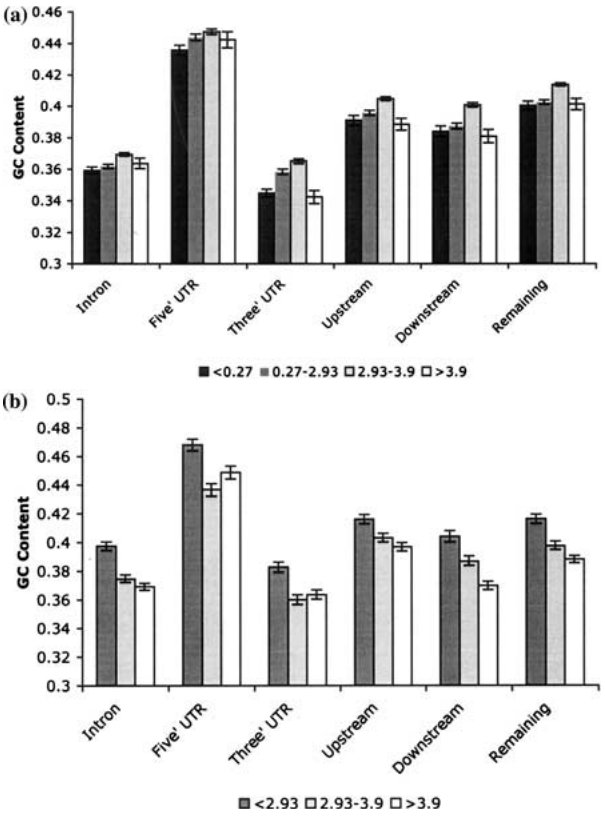

Fig. 2. Noncoding GC content on (**a**) the autosomes and (**b**) the X chromosome. Intron sequences are taken from genes whose total intronic sequence was greater than 200 bp in length. Five prime and three prime untranslated regions were included if their length exceeded 200 base pairs. Upstream and downstream sequences are 1000 bp flanking sequence surrounding transcribed sequences for genes that are sufficiently far from their nearest neighbor (see Materials and Methods for details). Remaining sequences are sequences more than 1 kb for genes in both directions. Error bars denote standard error.

noncoding regions, for instance, there is a consistent increase in noncoding GC content across the first three recombination categories, but there is also a conspicuous drop in noncoding GC content in the regions of highest recombination (Fig. 2). This could be due to a sampling effect; our recombination categories were constructed such that there were equal numbers of X-linked genes in each category (see Materials and Methods) and because recombination rates are higher on the X chromosome, there is a relative dearth of genes in this recombination category on the autosomes. Alternatively, the consistent reduction in GC content of noncoding DNA in regions of highest recombination on the autosomes could also reflect a non-linearity in this relationship.

On balance, it appears that there is an overall trend towards increased GC content with increased recombination on the autosomes, and decreased GC content with increased recombination on the X chromosome. This coupling between the base composition of coding and noncoding DNA may suggest that the forces that are modulating the base composition of coding sequences are operating similarly in noncoding sequences. Alternatively, the observation that the GC content of coding and noncoding sequences correlate independently with recombination rate might implicate distinct mechanisms for the maintenance of base composition in these two types of sequences. Lastly, the opposing directions of the relationships between GC content and recombination fate found the autosomes versus the X chromosome indicate that the forces governing base composition appear to be fundamentally different between the X chromosome and the autosomes.

Potential Explanations for the Negative Correlation Between GC Content and Recombination Rate on the X Chromosome

There are several hypotheses that could explain the observations that coding and noncoding GC content correlate negatively with recombination rate on the X chromosome. For instance, a background substitutional bias towards decreased GC with increased recombination rate could be mediated by a recombination-associated mutational bias towards A and T on the X chromosome. While hypothetically possible, this mutational explanation seems unlikely, as it would require a fundamentally different molecular mechanism of mutation specifically on the X chromosome, and to date, there are no empirical data to support such a difference in the underlying mechanisms of mutation between the chromosome sets.

Alternatively, our results could be explained by a background substitutional bias mediated by biased gene conversion. Given that gene conversion is believed to be generally GC-biased (Birdsell 2002), if rates of gene conversion on the X chromosome were highest in areas of low recombination, then this could lead to an X-specific, negative correlation between GC content and recombination rate. While it does seem possible that rates of gene conversion and rates of recombination are inversely related on the Drosophila X chromosome (Langley et al. 2000), this explanation would require an absence of such a relationship on the autosomes of the *D. melanogaster* genome.

The negative correlation between recombination rate and GC content on the X chromosome could also result from biases in background substitutional patterns resulting from selection at the level of global GC content. If the X chromosome had an optimal GC content of 30%, for instance, well below the autosomal average, and if the efficacy of selection on GC content were increased with increasing recombination due to Hill-Robertson effects, then a negative correlation between GC content and recombination rate would result. Were selection similarly maintaining GC content on the autosomes, this model would require a higher optimal GC content for autosomes

to explain the positive correlation between GC content and recombination rate.

Selection could also maintain this relationship between GC content and recombination rate by preserving sequence features intimately linked to genie function, such as functional regions within both coding sequences and noncoding sequences. If there were a bias in the base composition of functional sequences (coding and noncoding), and if this bias were different between the X and the autosomes, then selective pressure to preserve these compositionally biased regions could result in the systematic differences in recombination-associated patterns of base composition between the X and the autosomes.

It is further possible that the negative correlation between recombination rate and GC content reflects historical recombination rates rather than current ones, as it appears as if recombination maps are evolutionarily labile. Rates of recombination in *D. melanogaster*, for instance, appear to have decreased since the split from *D. simulans* (Takano-Shimizu 1999; True et al. 1996). One possible explanation for the negative correlations between recombination rate and the GC content of coding and noncoding sequences on the X and not the autosomes is that the X chromosome underwent a recent rearrangement of the recombination map such that the currently highly recombining areas experienced historically low recombination rates and visa versa.

Intuitively, if the recombination map on the X chromosome underwent such a dramatic rearrangement, then we might expect that there would be other differences between the X and the autosomes with respect to correlates of recombination rate. Interestingly, recombination rate also shows significant correlations with expression level on the X chromosome and the autosomes, and these correlations, too, are in opposite directions with a weak, negative correlation on the autosomes (Kendall partial correlation $\tau = -0.043$, Bonferroni-correlated $p = 0.02$) and a stronger, positive correlation on the X chromosome (Kendall partial correlation $\tau = 0.147$, Bonferroni-corrected $p << 0.001$) (Table 2). In addition, recombination rate and gene density are significantly and positively correlated on the autosomes, while no significant correlation between these two parameters exists on the X chromosome. It is not immediately clear why recombination rate should be correlated with gene density or expression level at all, but the lack of consistent correlations between recombination rate and other factors associated with genes on the autosomes and the X chromosome is curious and tenuously supports the supposition that recombination rates shifted on the X chromosome in the *D. melanogaster* lineage.

At this point, all of the above explanations appear contorted and none seems particularly likely. The parallel trends in GC content of coding and noncoding DNA on both the X chromosome and the autosomes may implicate biases in patterns of background substitution. However, the directions of these biases are reversed between the X chromosome and the autosomes, implicating X-specific effects. Scenarios invoking selection on functionally important sites in coding and noncoding regions can be envisioned as well, though there is no compelling evidence to date in favor of any individual hypothesis. Alternatively, the base composition of coding and noncoding DNA may be modulated by different forces, as partial correlation analysis suggests that the relationship between GC content of coding sequences and recombination rate is independent of the relationship between noncoding GC content and recombination rate and vice versa. It is clear that the unraveling of these effects will require much additional investigation.

Additional Relationships Among Codon Bias Correlates

In addition to the aforementioned correlations, our comprehensive analysis of codon usage in *D. melanogaster* confirmed some of the well-known relationships among genetic parameters on both the autosomes and the X chromosome and revealed several novel correlations as well. Here we report and briefly describe these relationships; a more complete treatment of these relationships will be presented elsewhere. Our data revealed several shared correlations between the X chromosome and the autosomes, with respect to both the magnitude of the correlation and the direction (Table 2). As previously reported (Bulmer 1988; Duret and Mouchiroud 1999; Hey and Kliman 2002; Sharp and Li 1986), codon bias is significantly positively correlated with expression level (Kendall partial correlation $\tau = 0.234$ and 0.221 for autosomes and X, respectively, Bonferroni-corrected $p << 0.0001$, both comparisons), which likely reflects increased selective benefits of translational efficiency for highly expressed genes. In addition, we confirm the negative correlation between protein length and codon bias (Kendall partial correlation $\tau = -0.315$ and -0.321 for autosomes and X, respectively. Bonferroni-corrected $p << 0.0001$, both comparisons). This may be because increased codon bias has stronger effects on translational efficiency in short genes as opposed to long genes (Akashi 1996; Duret and Mouchiroud 1999; Eyre-Walker 1996; Marais and Duret 2001) but could alternatively result from Hill-Robertson interference (Comeron et al. 1999).

We further confirm the strong negative correlation between codon bias and rate of protein evolution (Betancourt and Presgraves 2002; Schmid and Aquadro 2001) on both the autosomes and the X

chromosome (Kendall partial correlation $\tau = -0.431$ and -0.394 for autosomes and X, respectively, Bonferroni-corrected $p << 0.0001$, both comparisons) (Table 2). This relationship, too, may reflect different selective benefits of increased codon bias in genes evolving slowly rather than quickly. Relaxed selective constraints in rapidly evolving genes either at the level of codon-specific constraints (Akashi 1994) or gene-specific constraints (Comeron and Kreitman 1998) would result in the same pattern. Alternatively, because codon usage is thought to reflect weak selection on translational efficiency (Akashi 1995), this negative correlation between evolutionary rate and codon bias could reflect genetic hitchhiking, as the fixation probabilities of weakly selected mutations can be dramatically altered by strong directional selection at linked loci (Betancourt and Presgraves 2002; Kim 2004).

Our results also support the negative correlation between expression level and rate of protein evolution, which is found on both the autosomes and the X chromosome (Kendall partial correlation $\tau = -0.098$ and -0.107 for autosomes and X, respectively, Bonferroni-corrected $p \leq 0.02$, both comparisons) (Table 2). This is likely due to increased constraints on proteins expressed at high levels and/or in a broad range of tissues (Duret and Mouchiroud 2000; Krylov et al. 2003; Pal et al. 2001; Subramanian and Kumar 2004; Wright et al. 2004).

Gene density is positively correlated with optimal codon frequency on both the autosomes and the X chromosome (Kendall partial correlation $\tau = 0.051$ and 0.197 for autosomes and X, respectively, Bonferroni-corrected $p \leq 0.003$, both comparisons). This correlation has been previously found in the *D. melanogaster* genome (Hey and Kliman 2002), and the shape of the relationship appears to be non-linear, with the effect leveling off for the most densely packed genes (Hey and Kliman 2002). A Hill-Robertson interference model for selection on codon usage predicts that codon bias would be negatively correlated with gene density, which is precisely the opposite pattern from what is observed. Indeed, the cause of this positive correlation between gene density and codon bias is obscure.

In addition, on both the X chromosome and the autosomes, gene density and rate of protein evolution are significantly positively correlated (Kendall partial correlation $\tau = 0.068$ and 0.134 for autosomes and X, respectively. Bonferroni-corrected $p < 0.001$, both comparisons) (Table 2). One possible explanation for this association between gene density and rate of protein evolution is Hill-Robertson interference. Novel mutations in genes that are densely packed are more likely to interfere with one another's fixation probabilities than comparable mutations in genes with more intra and intergenic space (Comeron and Kreitman 2002). As a result, gene dense areas will be subject to weaker purifying selection; this reduction in the efficacy of selection in gene dense areas is akin to a relaxation of selective constraint in negatively selected genes, which may result in elevated rates of protein evolution. It is important to note, however, that gene density may not have been constant since the split of *D. melanogaster* and *D. pseudoobscura*, and as a result, the effect of gene density on rates of protein evolution may not be captured by our analysis.

To the extent that Hill-Robertson interference results in a positive correlation between gene density and rate of protein evolution, however, one might also expect a negative correlation between rate of protein evolution and recombination rate, but no such correlation is found on either the autosomes or the X chromosome. Again, this lack of correlation between rates of recombination and protein evolution could reflect changes in the recombinational environment in the *D. melanogaster* lineage since its split with *D. pseudoobscura*. The lack of concordance between the correlations of gene density and recombination with rate of protein evolution is thus an open question and merits further investigation.

Protein length and expression level are also significantly positively correlated on both the X chromosome and the autosomes (Kendall partial correlation $\tau = 0.255$ and 0.201 for autosomes and X, respectively, Bonferroni-corrected $p << 0.0001$, both comparisons) (Table 2). This may reflect ascertainment bias with respect to the expression data if for some reason the probability of finding an EST increases with protein length. Alternatively, this correlation could be indicative of a biologically relevant relationship between protein length and expression level. Much like the correlation between gene density and rate of protein evolution, this is an open question.

While many relationships among features associated with genes are shared between the X and the autosomes, our analyses also revealed differences between the X chromosome and the autosomes with respect to codon bias and its correlates (Table 2). For instance, the positive correlation between gene density and expression level on the autosomes (Kendall partial correlation $\tau = 0.108$, Bonferroni-corrected $p << 0.0001$) but not on the X chromosome (Kendall partial correlation $\tau = 0.043$, Bonferroni-corrected $p > 0.99$) (Table 2) was unexpected. Positive covariance between gene density and expression level may reflect increased selective benefits of gene clustering for highly expressed genes. This overall positive correlation between gene density and level of expression on the autosomes may actually result from a curvilinear relationship between these two features as reported by Hey and Kliman (2002), which may

reflect a tradeoff between the benefits to gene expression resulting from increased gene density and the cost of detrimental regulatory interactions as genes become more tightly packed (Hey and Kliman 2002). Independent of the shape of this relationship, however, the absence of this relationship on the X is difficult to explain.

In addition, there is a significant negative correlation between rate of protein evolution and protein length that is also unique to the autosomes (Kendall partial correlation $\tau = -0.111$, Bonferroni-corrected $p << 0.0001$) (Table 2). This may be due to Hill-Robertson interference; the efficacy of adaptive evolution on positively selected genes may be reduced with increased protein length because of the linkage to a larger number of functionally important sites in long genes versus short genes.

These differences between the X chromosome and the autosomes with respect to the relationships among genetic features may be due to the types of genes residing on these chromosomes. We do know, for instance, that there is a relative dearth of genes with male-biased expression on the X chromosome (Parisi et al. 2003). There may be other systematic differences between the gene complements of the X chromosome and the autosomes, which could explain the different relationships among genetic properties between the X chromosome and the autosomes.

Conclusions

Our comparison of codon usage patterns between the X chromosome and the autosomes of the *D. melanogaster* genome revealed a striking negative correlation between recombination rate and codon bias unique to the X chromosome. This is in contrast to the positive correlation between these two parameters on the autosomes. This negative correlation on the X chromosome is not a reflection of our measure of recombination rate, nor does it appear to result from an indirect correlation with known parameters. After controlling for all known correlates of codon bias, a strong and significant negative correlation between codon bias and recombination rate remains on the X chromosome. The base composition of noncoding DNA mirrors this trend as well, with GC content of noncoding DNA decreasing with increased recombination on the X chromosome. It is possible that background substitutional patterns are modulating the base composition of coding and noncoding sequences differently between the X chromosome and the autosomes, mediated by biased gene conversion, mutation, or selection on features other than the functional integrity of genes. It is similarly possible that the relationship between codon bias and recombination rate on the X chromosome is mediated by a yet unknown third variable.

Our results also reveal a number of interesting and unexpected relationships among properties of genes in the *D. melanogaster* genome. Not only are there systematic differences between the X chromosome and the autosomes that require convoluted explanations, but also, there are key similarities that are similarly puzzling. The large number of unexpected and difficult to explain findings highlights both the importance of treating the X chromosome and the autosomes separately and the complicated nature of the relationships among the properties of genes in the *D. melanogaster* genome.

Acknowledgments. This work was supported in part by the Stanford Genome Training Program (funded by 5 T32 HG00044 from the National Human Genome Research Institute) to N.D.S. and a Sloan Fellowship to D.A.P. Comments from two anonymous reviews, an associate editor, and the editor-in-chief considerably improved this manuscript.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136:927–935

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

Akashi H (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias. Faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144:1297–1307

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

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

Aquadro CF (1997) Insights into the evolutionary process from patterns of DNA sequence variability. Curr Opin Genet Dev 7:835–840

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

Betancourt AJ, Presgraves DC (2002) Linkage limits the power of natural selection in Drosophila. Proc Natl Acad Sci USA 99:13616–13620

Betran E, Thornton K, Long M (2002) Retroposed new genes out of the X in *Drosophila*. Genome Res 12:1854–1859

Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol 19:1181–1197

Bridges CB (1935) Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. J Hered 26:60–64

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection mutation balance? J Evol Biol 1:15–26

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–908

Caballero A (1995) On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics 139:1007–1011

Carvalho AB, Clark AG (1999) Intron size and natural selection. Nature (London) 401:344

Charlesworth B (1991) The evolution of sex chromosomes. Science (Washington DC) 251:1030–1033

Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. Am Natural 130:113–146

Comeron JM, Kreitman M (1998) The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: Mutation, selection or relaxed constraints? Genetics 150:767–775

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

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

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

Conery JS, Lynch M (2001) Nucleotide substitutions and the evolution of duplicate genes. Pacific Symposium on Biocomputing, pp 167–178

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

Duret L, 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

Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. Science (Washington DC) 303:537–540

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

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

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269–294

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

Kindahl EC (1994) Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*. Cornell University, Ithaca, NY

Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol 10:1239–1258

Kliman RM, Hey J (2003) Hill-Robertson interference in *Drosophila melanogaster*: Reply to Marais, Mouchiroud and Duret. Genet Res 81:89–90

Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:2229–2235

Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM (2000) Linkage disequilibria and the site frequency spectra in the su(s) and su(wa) regions of the *Drosophila melanogaster* X chromosome. Genetics 156:1837–1852

Laporte V, Charlesworth B (2002) Effective population size and population subdivision in demographically structured populations. Genetics 162:501–519

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

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

Marais G, Mouchiroud D, Duret L (2003) Neutral effect of recombination on base composition in *Drosophila*. Genet Res 81:79–87

Marin I, Siegal ML, Baker BS (2000) The evolution of dosage-compensation mechanisms. Bioessays 22:1106–1114

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B (2003) Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science (Washington DC) 299:697–700

Schmid KJ, Aquadro CF (2001) The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. Genetics 159:589–598

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Shields DC, Sharp PM, Higgins DG, Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Singh ND, Arndt PF, Petrov DA (2005) Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics 169:709–722

Sorsa V (1988) Chromosome maps of Drosophila. CRC Press, Boca Raton, FL

Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381

Takano-Shimizu T (1999) Local recombination and mutation effects on molecular evolution in *Drosophila*. Genetics 153:1285–1296

True JR, Mercer JM, Laurie CC (1996) Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics 142:507–523

Wright SI, Yau CBK, Looseley M, Meyers BC (2004) Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol 21:1719–1726

Yang Z (1997) PAML: am program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556