DOI: 10.1534/genetics.107.077081

Evolution of Gene Sequence in Response to Chromosomal Location

Carlos Díaz-Castillo[1] and Kent G. Golic

Department of Biology, University of Utah, Salt Lake City, Utah 84112

Manuscript received April 10, 2007
Accepted for publication June 6, 2007

ABSTRACT

Evolutionary forces acting on the repetitive DNA of heterochromatin are not constrained by the same considerations that apply to protein-coding genes. Consequently, such sequences are subject to rapid evolutionary change. By examining the *Troponin C* gene family of *Drosophila melanogaster*, which has euchromatic and heterochromatic members, we find that protein-coding genes also evolve in response to their chromosomal location. The heterochromatic members of the family show a reduced CG content and increased variation in DNA sequence. We show that the CG reduction applies broadly to the protein-coding sequences of genes located at the heterochromatin:euchromatin interface, with a very strong correlation between CG content and the distance from centric heterochromatin. We also observe a similar trend in the transition from telomeric heterochromatin to euchromatin. We propose that the methylation of DNA is one of the forces driving this sequence evolution.

DETAILED examination of the heterochromatic regions around eukaryotic centromeres has distinguished two subregions with differences in the structure of their chromatin and in their sequence composition (Heitz 1934; Gatti and Pimpinelli 1992). The central region, referred to as α-heterochromatin, hosts the centromere and is the most compacted chromosome region. In the polytene chromosomes of Drosophila, it shows the lowest degree of replication and consists mainly of highly repetitive elements. The region referred to as β-heterochromatin is generally thought to be located between the α-heterochromatin and the euchromatin of each chromosome arm. Its intermediate location also reflects the intermediate nature of its molecular and cytological characteristics. β-Heterochromatin is moderately compacted, moderately replicated in polytene chromosomes, and is formed by moderately repetitive elements interspersed with genes at a lower density than in euchromatic locations (Ashburner *et al.* 2004).

The β-heterochromatic genes present some unusual structural and regulatory characteristics. They span larger regions than is typical for euchromatic genes, mainly due to the possession of extremely large introns containing many insertions of transposable elements (Devlin *et al.* 1990; Biggs *et al.* 1994; Tulin *et al.* 2002; Dimitri *et al.* 2003). Although there is nothing obviously distinctive about the proteins that these genes encode, their regulation is in many ways contrary to that of euchromatic genes. Their expression is reduced when they are relocated away from centric heterochromatin (Khvostova 1939; Hessler 1958; Wakimoto and Hearn 1990; Eberl *et al.* 1993), and suppressors or enhancers of euchromatic gene variegation often have the opposite effect on the variegation of heterochromatic genes (Schultz 1936; Baker and Rein 1962; Wakimoto and Hearn 1990; Hearn *et al.* 1991; Lu *et al.* 2000; Weiler and Wakimoto 2002). Thus, both gene structure and expression appear to be influenced by a heterochromatic location.

Gene families are especially valuable for the study of molecular evolution since they afford the possibility of making several kinds of comparisons. One type of comparative analysis uniquely available with multi-gene families is the comparison of paralogs, those family members found within a single genome. Ideally, these analyses would make use of DNA sequence variation, both in coding and noncoding elements, gene exon structures and gene expression patterns, and known mutant phenotypes. These paralogous comparisons will help to detect conservation or divergence of function and/or structure and to deduce roles for each family member. Ultimately, this should allow us to interpret how such DNA and protein sequence changes are related to functional specializations, chromosome locations, or any other specific characteristic of the studied paralogs.

To specifically explore whether genes located near heterochromatin experience unique selective forces because of their location, we chose to examine the *Troponin C* (*TNC*) family of *Drosophila melanogaster*, which has members in euchromatin and in β-heterochromatin (Figure 1). TNC is the component of the sarcomeric thin filament that senses increases in cytosolic calcium

[1]*Corresponding author:* Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112.
E-mail: diazcastillo@biology.utah.edu

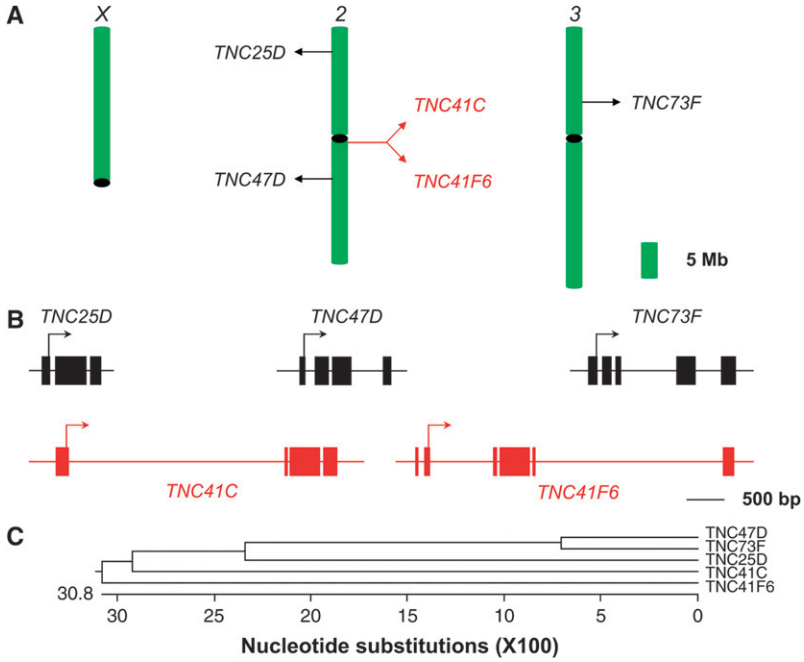

FIGURE 1.—(A) Chromosome locations, (B) exon structures, and (C) tree constructed aligning the nucleotide protein-coding sequences of the *TNC* genes of *D. melanogaster*. Heterochromatic genes are marked in red except in the tree.

and mediates myofibril contraction (FILATOV *et al.* 1999). Three of the five *TNC* genes of *D. melanogaster*, *TNC25D*, *TNC47D*, and *TNC73F*, are located in euchromatin and are expressed throughout development (HERRANZ *et al.* 2004). The genes that complete the family, *TNC41C* and *TNC41F6*, are located in the β-heterochromatin of *2R* and are expressed almost exclusively in late pupae and adults (HERRANZ *et al.* 2004). These β-heterochromatic members span larger chromosome regions than their euchromatic relatives due to the possession of larger introns (Figure 1). Since this is a common characteristic shared with other β-heterochromatic genes, it is probable that these genes have been present in a β-heterochromatin environment for some time and their characteristics may reveal the influence of such a chromosomal position.

It is true that it is somewhat unusual to use comparative analyses of paralogous genes because their origin and time of divergence may be very different, influencing their consideration as totally independent genes in the associated statistical analyses. We think this is not a big obstacle in this case. HERRANZ *et al.* (2005) estimated that the last duplication event affecting the *TNC* gene family in the *D. melanogaster* evolutionary line has occurred >60 MYA. It is known that much younger genes have accumulated a considerable amount of divergence (ZHANG *et al.* 2004), suggesting that the time passed since the last *TNC* duplication might have been more than enough for those genes to have accumulated changes independently. In fact, *TNC47D* and *TNC73F*, the two genes resulting from the last *TNC* duplication event that occurred in the *D. melanogaster* evolutionary line, diverged distinctively since the duplication, both with respect to their expression patterns and at the level of intron structure (HERRANZ *et al.* 2005).

The sequence characteristics of the *TNC* genes led us to look for similar tendencies in genes of *D. melanogaster* at all heterochromatin:euchromatin boundaries and within the predominantly or entirely heterochromatic chromosomes *4* and *Y*. Our findings are reported here.

MATERIALS AND METHODS

Sequences: The nucleotide sequences of the members of the *TNC* family of *D. melanogaster* were retrieved from FlyBase (FLYBASE 1999): *TNC25D* (Fbgn0031692), *TNC41C* (FBgn0013348), *TNC41F6* (FBgn0033027), *TNC47D* (FBgn0010423), and *TNC73F* (FBgn0010424). The data of genes located in the heterochromatin–euchromatin transition regions were retrieved from FlyBase (FLYBASE 1999) (Figures 4 and 5, and supplemental Data 2 at http://www.genetics.org/supplemental/).

To study the base composition of comparable tracts of the coding units located in the *Y* chromosome and their putative autosomal orthologs, we retrieved their nucleotide protein-coding sequences from public databases and used the annealing tools of Gene Jockey II Sequence Processor (Biosoft) to better define the regions that show the highest identity of sequence. The sequences of genes located in the *Y* chromosome were retrieved from the nucleotide databases of the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov): *kl-2* (AF313479), *kl-3* (AF313480), *kl-5* (AF136243), *ORY* (AF427496), *Ppl-Y1* (AF427493), *Ppl-Y2* (AF427494), and *Ppr-Y* (AF427495). The sequence of their putative autosomal paralogs were retrieved from FlyBase (FLYBASE 1999): *CG9068* (FBtr0087138), *CG9492* (FBtr0082100), *Dhc 93AB* (FBtr0084046), *CG6059* (FBtr0085130), *PpN58A* (FBtr0071734), *Ppl-87B* (FBtr0082595), *CG13125*-PA (FBtr0079898), and *CG13125*-PB (FBtr0079899).

Variation measures: The sequences of the *TNC* genes of *D. melanogaster* were initially aligned using Gene Jockey II

TABLE 1

Variation measures obtained from the comparison of the nucleotide protein-coding sequences of the *TNC* genes of *D. melanogaster*

	Euc. *vs.* Euc.			Het. *vs.* Euc. and Het. *vs.* Het.			
	TNC47D	*TNC73F*		*TNC25D*	*TNC41C*	*TNC47D*	*TNC73F*
			Total codon variation (S + N)				
TNC25D	0.64	0.62	*TNC41C*	0.81		0.75	0.78
TNC47D		0.33	*TNC41F6*	0.80	0.77	0.79	0.81
	Euc. *vs.* Euc.			Het. *vs.* Euc. and Het. *vs.* Het.			
	TNC47D	*TNC73F*		*TNC25D*	*TNC41C*	*TNC47D*	*TNC73F*
			Fraction of nonsynonymous codon varitation N/(S + N)				
TNC25D	0.72	0.74	*TNC41C*	0.58		0.52	0.50
TNC47D		0.30	*TNC41F6*	0.63	0.60	0.66	0.63
	Euc. *vs.* Euc.			Het. *vs.* Euc. and Het. *vs.* Het.			
	TNC47D	*TNC73F*		*TNC25D*	*TNC41C*	*TNC47D*	*TNC73F*
			Nonsynonymous codon variation (N)				
TNC25D	0.46	0.46	*TNC41C*	0.47		0.39	0.39
TNC47D		0.10	*TNC41F6*	0.50	0.46	0.52	0.51

	Euc. *vs.* Euc.		Het. *vs.* Euc. and Het. *vs.* Het.		
	Average	SD	Average	SD	*P*
		Statistical significance (Mann-Whitney test)			
S + N	0.53	0.17	0.79	0.02	0.017*
N/(S + N)	0.59	0.25	0.59	0.06	0.517
N	0.34	0.21	0.46	0.05	0.253

* Statistically significant; Het, heterochromatic *TNC* gene; Euc, euchromatic *TNC* gene. Heterochromatic genes are underlined.

Sequence Processor (Biosoft) and curated manually to minimize gaps and maximize the sequence identity (supplemental Data 1 at http://www.genetics.org/supplemental/). Since we were interested in the detection of evolutionary trends dependent on the relative location of the *TNC* genes, we based our measures on codon changes. Codon changes can reflect evolutionary trends at the protein level and at the genomic level.

In these comparisons, codon changes were divided into synonymous (S; codon changes that resulted in no change in the amino acid encoded) and nonsynonymous (N; codon changes that did produce a difference in the translated amino acid sequence) and were reported as the fraction of total codons. Changes that resulted in the deletion or insertion of codons were considered nonsynonymous. Multiple substitutions at a codon position were considered synonymous or nonsynonymous depending on the encoded amino acid regardless of the chain of substitutions that led to the current sequence.

Base composition and codon analyses: The base composition data of all protein-coding or noncoding nucleotide sequences were obtained using Gene Jockey II Sequence Processor (Biosoft) and ApE (http://www.biology.utah.edu/jorgensen/wayned/ape/).

The effective number of codons (N_c) (Wright 1990) of the *TNC* genes of *D. melanogaster* studied were calculated using CodonW 1.4.2 (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html). The N_c value is a measure of overall codon bias and ranges between 20 (when only 1 codon is used for each amino acid) and 61 (when codons are used randomly). GeneQuest 5.01 (DNASTAR) was used to obtain the codon frequency values used to calculate the data represented in Table 3 and Figure 3.

Statistical analyses: The statistical significance of the differences observed between the codon variation measures of two independent sets of *TNC* genes of *D. melanogaster* (Table 1) were analyzed using the nonparametric Mann-Whitney test provided by GraphPad InStat for Macintosh (GraphPad Software).

The statistical significance of the correlation of the changes in the codon frequencies partitioned according to whether they had been caused by transitions or transversions compatible with a decrease of the CG content in the third codon position of heterochromatic *TNC* genes (Table 3 and Figure 3) was analyzed using the nonparametric Spearman test provided by GraphPad InStat for Macintosh (GraphPad Software).

The statistical significance of the correlation between genes' protein-coding CG composition and distance from centromere or telomere (Table 5, Figures 4 and 5, and supplemental Data 2 at http://www.genetics.org/supplemental/) was analyzed using the nonparametric Spearman test provided by GraphPad InStat for Macintosh (GraphPad Software).

Other software used: Power Macintosh MegAlign 5.01 (DNASTAR) was used to draw a phylogenetic tree of the *TNC* genes of *D. melanogaster* based in the alignments of the nucleotide protein-coding sequences used in the codon variation analyses (Figure 1 and supplemental Data 1 at http://www.genetics.org/supplemental/). Microsoft Excel X for Mac was used to manipulate and represent data. Microsoft PowerPoint X for Mac, Microsoft Word X for Mac, and Adobe

Illustrator 10 were used in the preparation of the manuscript, tables, and figures.

RESULTS

Comparison of euchromatic and heterochromatic *TNC* genes: The examination of the genomes of related species has revealed that heterochromatic DNA can change rapidly during evolution (Powell 1997). More specifically, the centromeric DNA of eukaryotic species shows rapid evolution (Henikoff *et al.* 2001). Although this high rate of change in heterochromatin pertains to the noncoding sequences that are located in α-heterochromatin, we wondered whether a similar trend might be visible in the genes located in the adjacent β-heterochromatin. A comparative study of a paralogous group of genes with euchromatic and heterochromatic locations, like those that encode for TNC in *D. melanogaster*, might be a good way to answer this question (Figure 1).

If the heterochromatic *TNC* genes are under a rapid evolution trend, similar to the one reported for the centromere repeats or heterochromatic DNA in general, we should be able to detect it at the nucleotide level and possibly even at the protein level. An examination of the codons of these genes allows detection of both nucleotide and the protein variation. We compared the protein-coding sequences of the *TNC* genes two at a time, quantified the divergent codons in each of those comparisons, and presented the results as a fraction of the total number of codons compared.

Synonymous changes (S) are those in which the divergence detected between two codons does not result in an amino acid change; nonsynonymous changes (N) are those that do produce an amino acid change. Changes that resulted in the partial or total deletion or insertion of codons were considered nonsynonymous. The addition of synonymous and nonsynonymous changes (S + N) will inform us of the degree of variation accumulated at the nucleotide level, while nonsynonymous changes either in absolute (N) or relative [N/(S + N)] terms will tell us more specifically about the variability that has occurred at the protein level.

A special problem when dealing with the analysis of variability at the codon level is posed by multiple substitutions in a single codon. In such cases, there is no easy way of knowing if those were caused by a succession of only one kind of substitution or by the alternation of synonymous and nonsynonymous substitutions. Some approaches have been devised to deal with this inconvenience (for general reference, use Graur and Li 2000). These approaches make *a priori* considerations about the probability of the different base substitutions: either the probability of the individual base changes is the same in all cases or they aren't and we need to introduce corrections to somehow reflect the biases (for instance, usually synonymous substitutions are more frequent than nonsynonymous). At this point in our study, we chose not to make assumptions about what type of sequence substitutions might occur in heterochromatic environments, because the eventual detection of biases at this level is the final aim of our work. Therefore, we limited our consideration to synonymous or nonsynonymous changes, regardless of the chain of events that drove to these changes. This approach ensures that we are not introducing biased assumptions at the outset.

Once we obtained the values for synonymous and nonsynonymous variation in the *TNC* genes, we partitioned them into two groups. One group is formed by the parameters obtained from comparisons in which at least one of the genes is heterochromatic. These data will inform us of the degree of variation heterochromatic *TNC* genes accumulate on average. The other group is formed by the parameters obtained from comparisons of two euchromatic genes. This second group will act as a baseline to allow us to determine if the heterochromatic *TNC* genes differ significantly in their accumulated changes.

We found that the heterochromatic *TNC* genes have accumulated higher codon divergence (S + N) than the euchromatic genes (Table 1; $P = 0.0167$). However, this has not led to an increased variation in the proteins that the heterochromatic genes encode. When comparing the average divergence of euchromatic genes with the average divergence of heterochromatic genes, the fraction of codon changes that produce changes in the amino acid sequence [N/(S + N)] does not differ significantly (Table 1; $P = 0.5167$). Even when considered as the absolute number of amino acid changes (N), the heterochromatic genes do not show a significantly increased degree of variation (Table 1; $P = 0.2526$).

Thus, heterochromatic *TNC* genes seem to accumulate elevated nucleotide variation, although this variation is not translated into an elevated variation of the proteins they encode. This result suggests that the proteins encoded by the heterochromatic genes are subject to the same functional constraints as the euchromatic genes, consistent with their role as the primary sources of TNC in the adult muscles (Herranz *et al.* 2004). The elevated nucleotide variation is reminiscent of the rapid evolution detected for noncoding elements of heterochromatic centromeric elements (Henikoff *et al.* 2001).

Specific biases detected in heterochromatic *TNC* genes: To try and gain some understanding of the possible causes of the elevated nucleotide variation of the heterochromatic genes, we analyzed the sequence with respect to other parameters, such as base composition and codon bias. The initial analysis of the *D. melanogaster* genome sequence revealed that the DNA at the base of each chromosome arm, the β-heterochromatic region, had a reduced CG content relative to the euchromatic portions of each arm (Adams *et al.* 2000). However, it wasn't determined whether the decreased CG content was a result of interspersed AT-rich repetitive sequences,

TABLE 2

Base composition, CG content, and N_c values of the nucleotide protein-coding sequences of the *TNC* genes of *D. melanogaster*

	TNC25D	*TNC41C*	*TNC41F6*	*TNC47D*	*TNC73F*
A%	27	30	32	25	24
C%	26	17	17	26	25
G%	29	25	25	29	31
T%	19	27	26	20	19
CG%	55	42	42	55	56
N_c	42.52	49.03	61.00	35.88	31.46

Heterochromatic genes are underlined.

like those found in the adjacent α-heterochromatin, or whether the protein-coding segments also had a reduced CG content. To test whether a general reduction in CG content might account for the increased variation of the heterochromatic *TNC* genes, we determined the base composition of the five *TNC* genes and found that the β-heterochromatic genes are CG-depleted relative to their euchromatic counterparts (Table 2). This reduction affects both protein-coding and non-protein-coding sequences (Figure 2). This finding that the β-heterochromatic *TNC* genes exhibit a reduced CG content in their coding regions clearly shows that this is not solely due to the accumulation of repetitive sequence elements in the noncoding parts of the gene and further suggests that it is not merely a regulatory adaptation specific for the region.

Finally, we also detected a difference between the heterochromatic and the euchromatic *TNC* genes of *D. melanogaster* for N_c (Wright 1990). N_c values are higher in the case of heterochromatic *TNC* genes (Table 2), which is an indication that those genes have lower codon bias than their euchromatic relatives.

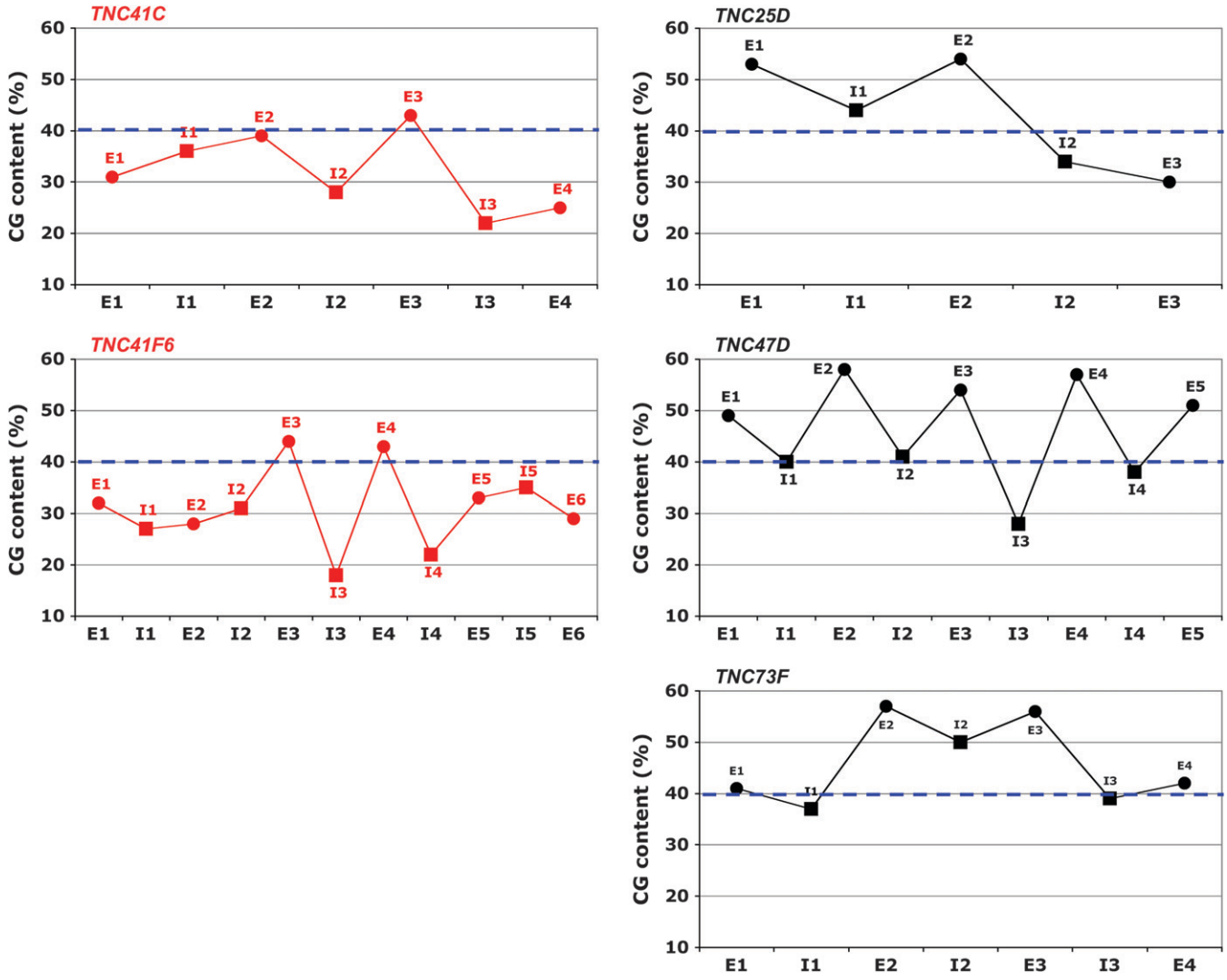

Figure 2.—CG content (%) of exons (E, circles) and introns (I, squares) of the *TNC* genes of *D. melanogaster*. A blue dashed line has been arbitrarily placed at the level of the 40% of CG content to better compare the five graphs. Heterochromatic genes are marked in red.

In summary, the heterochromatic *TNC* genes of *D. melanogaster* have elevated nucleotide variation relative to the euchromatic paralogs; even though the encoded proteins do not vary significantly from those encoded by the euchromatic genes, they exhibit a depletion of the CG content of coding and noncoding elements and they have reduced codon biases.

Biased mutation as a mechanism for heterochromatic CG depletion: The fact that both protein-coding and noncoding sequences of the heterochromatic *TNC* genes of *D. melanogaster* exhibit CG depletion without a significant alteration in the proteins they encode suggests that this cannot be explained by a selection bias. Instead, we considered whether some force acting directly at the level of DNA could be responsible for the reduced CG content of the heterochromatic genes. In other words, we wondered whether a biased mutational process could account for the differences between the heterochromatic and euchromatic genes. For example, as we will later discuss, it has been proposed that the occurrence of meiotic recombination may exert a mutational bias toward CG in the euchromatic regions where recombination occurs (Marais 2003). We wondered whether there might be evidence for a biased mutational process in heterochromatic regions. By comparing the different extents of common trends shown by the heterochromatic *TNC* genes of *D. melanogaster*, we might be able to identify traces of such a mechanism. Therefore, we undertook an analysis of some of the sequence parameters obtained from the *TNC* genes of *D. melanogaster*.

We studied the base composition of the three codon positions by analyzing the nucleotide protein-coding sequences of the *TNC* genes of *D. melanogaster*. The average CG percentages found in the heterochromatic *TNC* genes of *D. melanogaster* are: 1st = 58.8, 2nd = 30.6, and 3rd = 37.8. For the euchromatic *TNC* genes of *D. melanogaster*, the values are: 1st = 60.5, 2nd = 27, and 3rd = 78.2. The differences in CG content for the heterochromatic genes at each position are: 1st = −1.7, 2nd = 3.6, and 3rd = −40.4, indicating that the drop in CG content detected in the heterochromatic *TNC* genes of *D. melanogaster* is based almost entirely on CG depletion of the third codon position, whereas the composition of the first and second positions shows very little change. This is not a surprise considering the degeneracy of the genetic code and is consistent with our finding that the proteins encoded by these genes have not significantly diverged.

Next, we examined the frequencies of codons grouped according to the composition of the second and the third positions. The practical immutability of the bases in the second codon position allowed us to study possible trends in the changes that occurred at the third codon position, both at the single nucleotide and the dinucleotide level. If a mutagenic process were to work mainly to reduce C and G bases in the third codon positions, we expect the frequency of codons with C or G in the third position in heterochromatic genes to decrease, while the frequency of codons with T or A in the third position should increase. Based on the results we presented in the Table 3, we can say that this is generally correct for the heterochromatic *TNC* genes of *D. melanogaster*. With the exception of NGG and NCG that show a slight increase of frequency in heterochromatic genes, the frequencies of codons ending in T or A are increased, whereas the frequencies of the other codons ending in C or G are decreased in heterochromatic genes.

TABLE 3

Changes of the frequencies of codons group according to the composition of their second and third bases in the heterochromatic *TNC* genes of *D. melanogaster*

Codon	Frequency difference[a]	Codon	Frequency difference[a]
NAA	0.1029	NGG	0.0044
NTT	0.0813	NCG	0.0029
NCT	0.0476	NAC	−0.0245
NTA	0.0435	NGC	−0.0351
NAT	0.0432	NTG	−0.0406
NGA	0.0347	NCC	−0.0696
NCA	0.0325	NTC	−0.0926
NGT	0.0183	NAG	−0.1487

[a] Value calculated by subtracting the average frequency of each codon group in the euchromatic *TNC* genes from the average frequency of each codon group in the heterochromatic *TNC* genes of *D. melanogaster*.

Mutations that reduce CG content can be of two kinds: transitional (C to T and G to A) or transversional (C to A and G to T). We wondered to what degree the third position CG depletion could be attributed to each of these two types of mutation. To determine that, we partitioned the changes of the codon frequencies found in Table 3 according to whether they had been caused by transitions or transversions in the third position that resulted in a decrease of the CG content. For instance, the decrease of CG content of a protein-coding sequence based in transitions occurred in the third base of NAC codons will result in NAT, whereas the transversion will result in NAA. This arrangement of the data is plotted in Figure 3, showing that in *D. melanogaster*, there is a statistically significant negative correlation when arranged according to transitions ($r = -0.8333$; $P = 0.0154$) but not according to transversions ($r = -0.2381$; $P = 0.9768$). This means that the codons of a pair were enriched or depleted in comparable magnitudes only when the third position changes were transitions. Thus, the reduction in CG content of the heterochromatic *TNC* genes of *D. melanogaster* might have been caused by the action of a mutagenic mechanism that mainly generates transitions.

If the main mechanism responsible for CG depletion of the heterochromatic *TNC* genes of *D. melanogaster* is indeed based in the production of transitions in the third codon position, we should most easily detect this

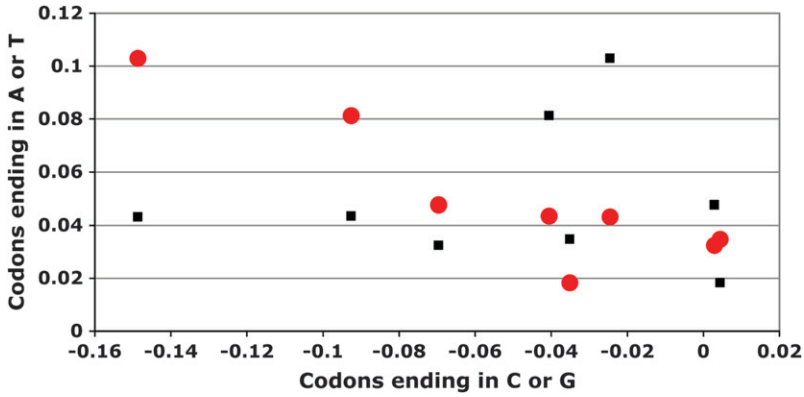

FIGURE 3.—Graph representing the data from Table 3 partitioned according to transitions (red circles) or transversions (black squares) that occurred in third codon bases would result in a decrease of the CG content of the protein-coding sequences of heterochromatic *TNC* genes. The data of codons ending with C or G, supposedly depleted, are represented in the *x*-axis, and the data of the codons ending with A or T, supposedly enriched, are represented in the *y*-axis.

trend when analyzing the codons that specify amino acids based on only on the first two positions: serine (UCN), leucine (CUN), proline (CCN), arginine (CGN), threonine (ACN), valine (GUN), alanine (GCN), and glycine (GGN). This time, we compared the frequency of changes that occurred in the third base of this subset of codons in two groups of protein-coding sequence alignments of the *TNC* genes of *D. melanogaster*, which already permitted us to obtain the codon variation parameters (supplemental Data 1 at http://www.genetics.org/supplemental/). The first set of alignments was formed by comparing one heterochromatic *TNC* gene and one euchromatic *TNC* gene, whereas the second set of alignments was formed by comparing two genes that were euchromatic. When we subtract the average frequencies of third base codon changes obtained in the euchromatic *TNC* gene *vs.* euchromatic *TNC* gene comparisons from the average values obtained in the heterochromatic *TNC* gene *vs.* euchromatic *TNC* gene comparisons, we will have an indication of the type of changes that are preferred in the heterochromatic *TNC* genes. As expected, in Table 4, we can see that changes that decrease CG content are elevated in the heterochromatic *TNC* genes, whereas those that would increase CG content in the third codon position are reduced. In the case of changes that result in loss of C alone, the transition (NNC to NNT) is more frequent than the transversion (NNC to NNA). However, the opposite is true for changes that result in the depletion of G; transversions (NNG to NNT) are more common than transitions (NNG to NNA).

DNA methylation as a mechanism for CG reduction: The methylation of cytosine increases its tendency to spontaneously deaminate (SHEN *et al.* 1994). While the unmodified base is deaminated into uracil, methylated cytosines deaminate into thymines. Methylated sequences of very different organisms are known to be hotspots for sequence variation and are characteristically enriched in C to T and G to A transitions (COULONDRE *et al.* 1978; COOPER and YOUSSOUFFIAN 1988; SELKER 1990; JONES *et al.* 1991; GREENBLATT *et al.* 1994; SINGER *et al.* 1995; YANG *et al.* 1996; COLOT ET ROSSIGNOL 1999; WATTERS *et al.* 1999). Moreover, the methylation of DNA has been proposed to contribute to regional differences in base composition found in other genomes (FRYXELL and ZUCKERKANDL 2000; EYRE-WALKER and HURST 2001).

If DNA methylation was responsible for the trends we found in the heterochromatic *TNC* genes of *D. melanogaster*, we should find noticeable traces of its activity in the analyses we just presented. One of those expected traces would be the depletion of CT dinucleotides, since CT is reported to be the preferentially methylated dinucleotide in *D. melanogaster* (LYKO *et al.* 2000). The two codons produced from DNA with a CT dinucleotide in the last position are either NCT or NAG (with CT on the complementary strand). Of these, NAG is more amenable to change, since deamination, and a C to T transition on the opposite strand, will produce no change in the encoded amino acid (NAG to NAA). As shown in Table 3, NAG codons show the greatest decrease and NAA the greatest increase in the heterochromatic *TNC* genes of *D. melanogaster*.

TABLE 4

Changes of the frequencies of third base mutations in codons that specify amino acids based in their two first bases in the heterochromatic *TNC* genes of *D. melanogaster*

Mutation	Frequency difference[a]	Mutation	Frequency difference[a]
NNC to NNT	0.1482	NNA to NNG	−0.0105
NNC to NNA	0.0837	NNT to NNC	−0.0523
NNT to NNA	0.0744	NNC to NNG	−0.0649
NNG to NNT	0.0730	NNA to NNC	−0.0780
NNG to NNA	0.0389	NNT to NNG	−0.0853
NNA to NNT	0.0195	NNG to NNC	−0.1469

[a] Value calculated by subtracting the average frequency of each mutation obtained from the alignments of the nucleotide protein-coding sequences of two euchromatic *TNC* genes from the average frequency of each mutation obtained from the alignments of the nucleotide protein-coding sequences of one heterochromatic and one euchromatic *TNC* gene.

Our results also show that the frequency of codons ending with CT increases, a finding that does not follow directly from the hypothesis that methylation of C in CT dinucleotides is responsible for CG depletion of heterochromatic genes. However, the decrease of CG content in the heterochromatic *TNC* genes of *D. melanogaster* is almost entirely the result of changes in the third codon base, whereas the second position barely changes. As discussed previously, this is likely a result of selection for conservation of protein function. Therefore, it might only be possible to increase the frequency of codons ending with CT as a result of transitions that occur in the third position of codons ending with CC. Thus, the sequence differences between heterochromatic and euchromatic *TNC* genes are compatible with changes provoked by cytosine methylation.

If a DNA methylation-based mechanism is responsible for the CG depletion at the third codon position, we should also find a higher frequency of CG-reducing transitions in those codons that specify amino acids based on the first two positions. Consistent with this expectation, NNC to NNT substitutions are by far the most frequent changes in the heterochromatic *TNC* genes of *D. melanogaster* (Table 4). Quite surprising is the result that NNG to NNA changes, though increased, do not show a higher frequency in the same set of genes. Furthermore, contrary to our expectations, NNG to NNT transversions seem to be clearly preferred to NNG to NNA transitions in the heterochromatic *TNC* genes of *D. melanogaster*. Though the elevated frequency of NNC to NNT substitutions speaks in favor of the activity of a DNA methylation-based mechanism, the somehow lower-than-expected frequency of NNG to NNA transitions might indicate that this is not the only mechanism responsible for the CG depletion found in the heterochromatic *TNC* genes of *D. melanogaster*.

CG content of protein-coding segments in heterochromatic-euchromatic transition regions: The data presented above support the possible existence of a DNA methylation-based mechanism that, at least in part, has caused the accumulation of AT-biased sequence variation in the heterochromatic *TNC* genes of *D. melanogaster*. If the CG depletion is caused by methylation of heterochromatic DNA, then it seems probable that the AT composition bias should not be restricted to the heterochromatic *TNC* genes but should be seen with all heterochromatic genes. To determine whether the reduced CG content of *TNC41C* and *TNC41F6* reflects a regional influence, we examined the protein-coding regions of genes found in the first ~3 Mbp of DNA sequence at the base of *2R* in the vicinity of both genes. The available sequence at the base of each chromosome arm begins where the repetitive content is reduced sufficiently to allow the assembly of shotgun-sequenced fragments—essentially, in β-heterochromatin. We extended our analyses far enough from the centromere to be sure that some of the genes studied were clearly located in euchromatin. The result, presented in Figure 4, shows that the CG depletion of gene protein-coding sequences in β-heterochromatin of *2R* is characteristic of that region as a whole, and the strength of the effect depends on the proximity to α-heterochromatin.

Because the reduction in CG content was not specific to the *TNC* genes, we became interested in whether CG reduction was a common characteristic of the protein-coding sequences of genes in the vicinity of heterochromatin. We surveyed genes located in all heterochromatin–euchromatin transition regions of the genome. Our survey encompassed genes found near centromeric and telomeric heterochromatin. Most of these regions on the major chromosomes (*X*, *2*, *3*) appear to show trends of a general increase in CG content of protein-coding segments as the genes become more distant from the nearest heterochromatic region (Figures 4 and 5). Around centromeres, the CG content is higher for genes that are more distant from the centromere (Figure 4). At the chromosome tips, CG content is higher for genes with a more internal location, placing them further from the telomere (Figure 5). Those general trends are manifest to differing extents in the surveyed regions, and in 8 of the 10 surveyed regions on *X*, *2*, and *3*, the trend was statistically significant (Table 5).

The strength of the correlations between CG content and distance to the centromere or telomere are not homogeneous, as the *r* values compiled in Table 5 show. Some of these *r* values are rather small, meaning that many genes in these regions do not strictly conform to the CG trend. This suggests the rather unsurprising conclusion that other forces may also act to influence the DNA sequence of genes in these regions.

Within the genome of *D. melanogaster*, the chromosomes with the highest proportion of heterochromatin by far are the *Y* and *4* (Gatti and Pimpinelli 1992). If the reduction in CG content is truly a result of a heterochromatic environment, we should notice it very clearly in these two chromosomes. The tiny chromosome *4* is known to have characteristics of heterochromatin throughout its length (Sun *et al.* 2000; Sun *et al.* 2004). Accordingly, we find that the CG contents of the coding regions of almost all the genes along this chromosome are very low (Figure 4). Despite that, no statistical significance was found when studying the correlation of the CG contents and the distance to the centromere of the chromosome *4* (Table 5). Wang *et al.* (2002), who surveyed nucleotide variation in a worldwide sampling of chromosomes *4* of *D. melanogaster*, identified a region, limited by the genes *CG11091* and *toy*, with levels of variation typical of other putatively euchromatic autosomal regions. Visual inspection of the data presented in Figure 4 does seem to indicate a region of somewhat higher CG content coinciding with the *CG11091-toy* region. If the most polymorphic region of chromosome *4* can be taken as an indication of its most

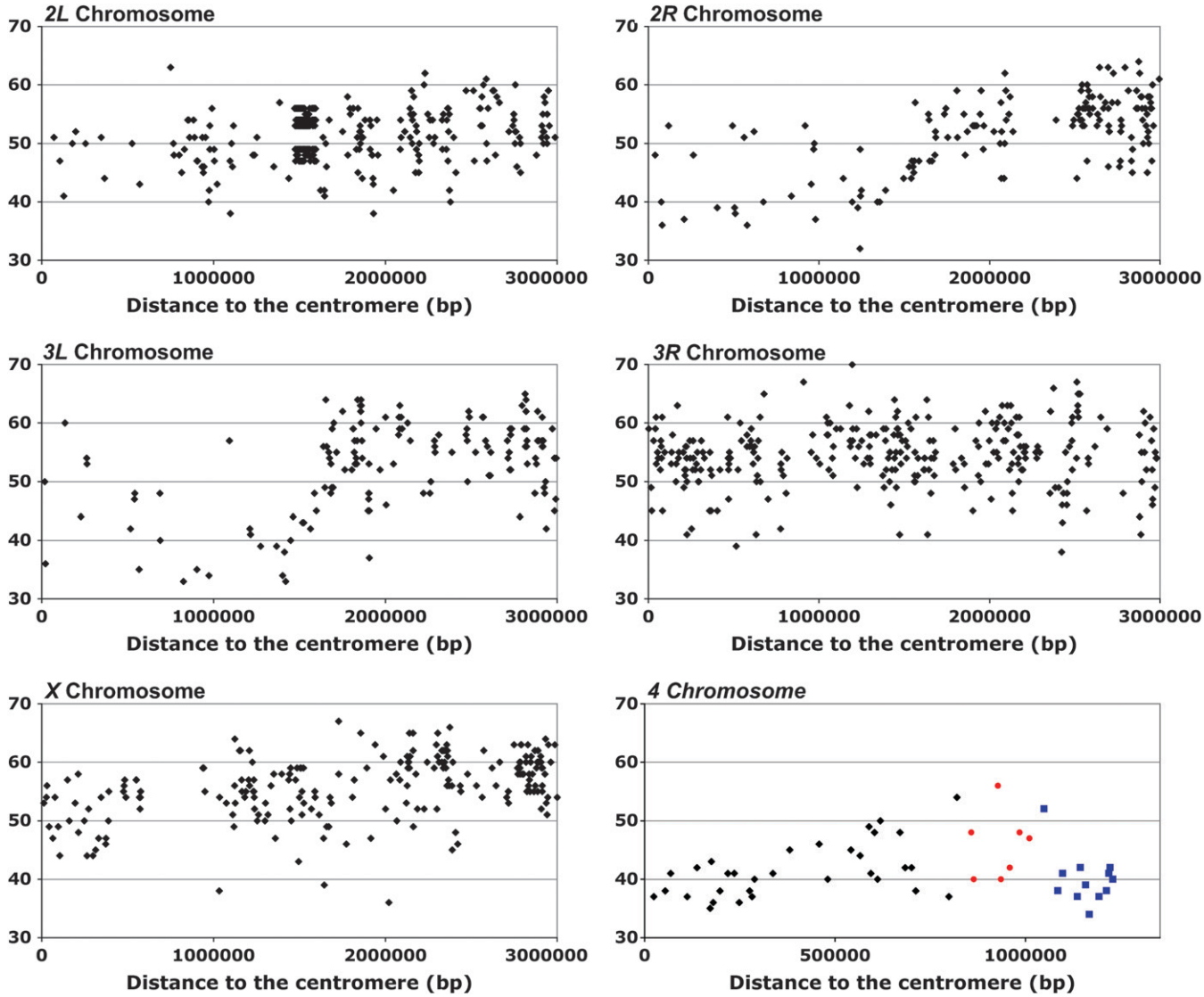

FIGURE 4.—CG content (%) of the nucleotide protein-coding sequences of genes located in the main heterochromatin–euchromatin transition regions of centromeres and the chromosome *4* of *D. melanogaster*. The data are always represented with the heterochromatic side of each region toward the left side of the graphs and the euchromatic side toward the right side of the graphs regardless of the real orientation they have in *D. melanogaster* chromosomes. Different symbols were used to distinguish three regions of the *fourth* chromosome (see main text for further explanations).

euchromatin-like region, the relevance of the data of the chromosome *4* is even higher. Replicating the analyses of the other heterochromatin–euchromatin transition regions, we studied the statistical significance of the correlation of the CG contents of the protein-coding sequence of the genes located in chromosome *4* and their distance to its centromere going as far as to include genes presumably located in euchromatin, those within the region *CG1109-toy*. We considered the region located between the centromere and the gene *toy* as the centromeric transition region of chromosome *4* and the region between the gene *CG11152* and the telomere as the telomeric transition region. With this assumption, we found that the correlation between CG content of protein-coding sequences and their distance from the centromere or telomere was also statistically significant (Table 5).

Finally, a further demonstration of the trend for heterochromatic regions to exhibit reduced CG content comes from analysis of the *Y*. This chromosome is considered to be totally heterochromatic (GATTI and PIMPINELLI 1992). Consistent with that, the number of coding elements is very reduced (GATTI and PIMPINELLI 1992; ASHBURNER *et al.* 2004). Most of the closest paralogs of single-copy *Y*-linked genes are found in autosomes (CARVALHO *et al.* 2001). When we analyzed the CG content of coding units located in the *Y* and their putative autosomal paralogs, we saw that the former were clearly CG-depleted (Table 6). In the case of the genes located in the *Y* chromosome, it is hard to test if

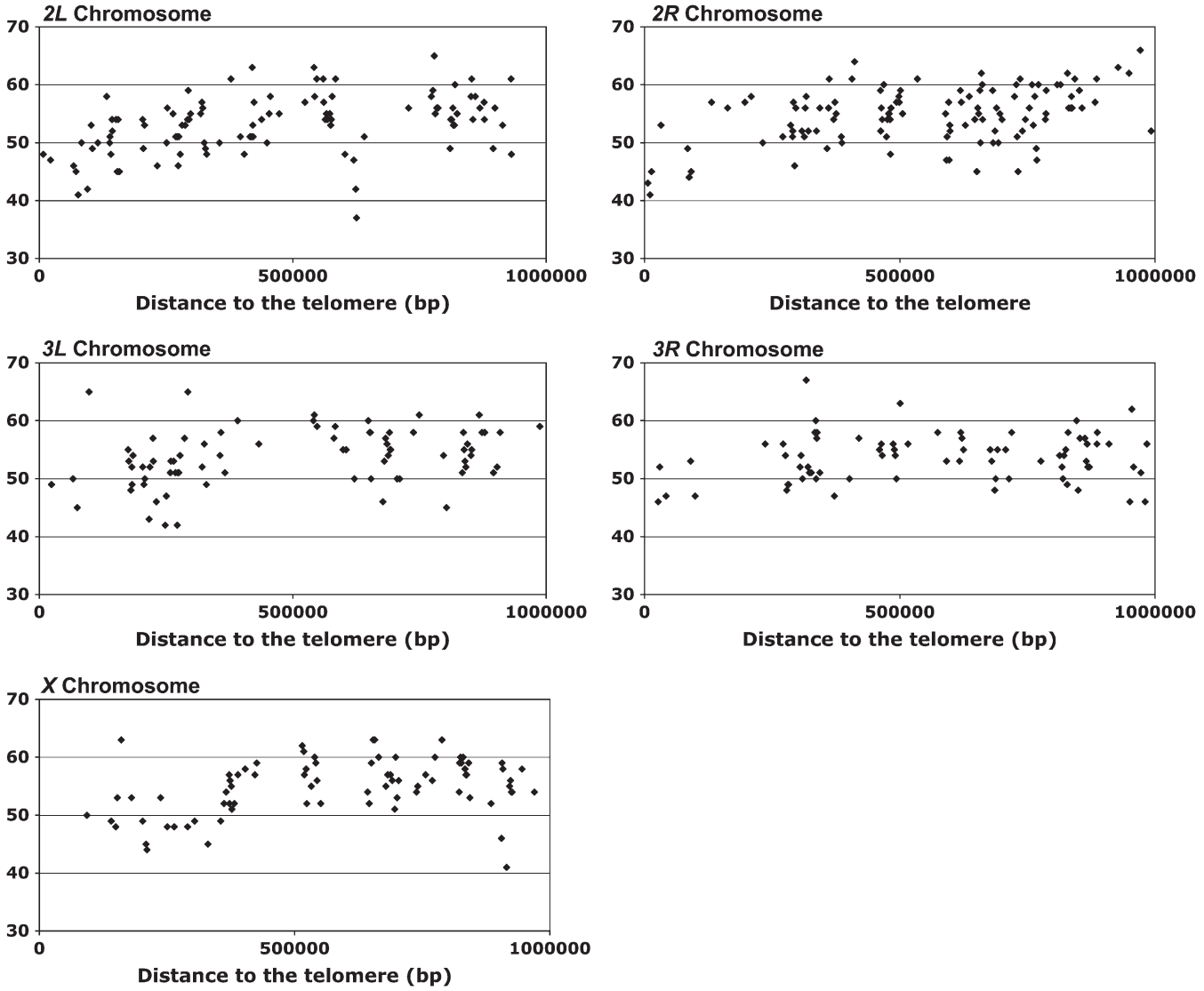

FIGURE 5.—CG content (%) of the nucleotide protein-coding sequences of genes located in the main heterochromatin–euchromatin transition regions of telomeres of *D. melanogaster*. The data is always represented with the heterochromatic side of each region toward the left side of the graphs and the euchromatic side toward the right side of the graphs regardless of the real orientation they have in *D. melanogaster* chromosomes.

the proximity to telomeres or centromeres affects the degree of the CG depletion, since many of these coding units hadn't been finely mapped. It is true though that the CG content of *kl-2*, *kl-3*, and *kl-5* progressively increases as their distance from the centromere increases (Table 6) (GATTI and PIMPINELLI 1992).

DISCUSSION

It has been previously noted that a gradient of increasing CG content is apparent in the chromosomal DNA of *D. melanogaster* as it passes from centric heterochromatin to euchromatin (ADAMS *et al.* 2000). By specifically examining the protein-coding regions of genes in these parts of the chromosomes, we found that this result cannot be solely explained by the accumulation of transposable elements of low CG content or as an adaptation of noncoding DNA to a heterochromatic location. Our analysis of genes at the boundaries of heterochromatin and euchromatin reveals that the composition of the protein-coding sequences of genes in such locations reflects their position. Genes closer to centric or telomeric heterochromatin tend to have a lower CG content than genes that are more removed, and this tendency is proportional to their proximity to heterochromatin.

The *TNC* gene family has members located in euchromatin and in heterochromatin. Our analysis shows that, though the proteins encoded by the heterochromatic genes are not more divergent than the euchromatic genes, the underlying DNA sequence has diverged significantly, with the heterochromatic family members

TABLE 5

Statistical analyses of the CG content patterns in the main heterochromatin–euchromatin transition regions and the chromosome *4* of the genome of *D. melanogaster* represented in Figure 4

Chromosome compartment	No. of protein-coding genes studied	Total no. of protein-coding genes	Coverage (%)	Spearman test (95% confidence interval)	
				r	*P*
*X*c	212	258	82	0.4458	<0.0001*
*X*t	79	94	84	0.3617	0.0011*
*2L*c	285	331	86	0.2162	0.0002*
*2L*t	105	127	83	0.4437	<0.0001*
*2R*c	156	192	81	0.5766	<0.0001*
*2R*t	115	139	83	0.3728	<0.0001*
*3L*c	144	167	86	0.3775	<0.0001*
*3L*t	80	106	75	0.3704	0.0007*
*3R*c	297	374	79	0.0966	0.0966
*3R*t	75	92	82	0.1597	0.1710
4	51	90	57	0.1511	0.2900
*4*c	39	70	56	0.5179	0.0007*
*4*t	19	35	54	−0.4771	0.0389*

* Statistically significant; c, centromere; t, telomere.

showing a much lower CG content. This example strongly suggests that the DNA sequences of the genes located in heterochromatin, while still constrained by the necessity of encoding functional proteins, are evolving in response to the influence of their chromosomal location.

The evolution of heterochromatin, and of genes in the vicinity of heterochromatin, has interested geneticists since the discoveries that heterochromatin exhibited many unique properties, including housing the sites for chromosome segregation (Anderson 1925), having novel modes of gene regulation (Muller and Painter 1932), and possessing very distinct sequence content (Kliman and Hey 1993; Hey and Kliman 2002; Ashburner *et al.* 2004). Regional differences of CG content have been detected within the genomes of several organisms (Bernardi 1989; Sharp *et al.* 1989; Carulli *et al.* 1993; Sharp and Lloyd 1993; Dujon *et al.* 1994; Feldmann *et al.* 1994; Bernardi 1995; Deschavanne and Filipski 1995; Bradnam *et al.* 1999; Eyre-Walker and Hurst 2001; Daubin and Perrière 2003; Zhang and Zhang 2004), and a positive correlation between CG content and recombination rates has been identified (Ikemura and Wada 1991; Gerton *et al.* 2000; Fullerton *et al.* 2001; Marais *et al.* 2001, 2003; Birdsell 2002; Kong *et al.* 2002; Marais and Piganeau 2002; Meunier and Duret 2004). It has long been known that heterochromatic regions have greatly reduced levels of recombination (Kliman and Hey 1993; Hey and Kliman 2002; Ashburner *et al.* 2004). The basis for the correlation between reduced recombination and CG content has been the subject of much interest.

One class of model supposes that the reduced CG content of heterochromatic regions is an indirect consequence of the reduction in recombination in these

TABLE 6

CG content of comparable regions of some of the coding units found in the chromosome *Y* of *D. melanogaster* and their putative orthologs located out of this chromosome

Y paralogs			Autosomal paralogs		
Genes	Sequence fragment (bp)	CG %	Genes	Sequences fragment (bp)	CG (%)
kl-2	2-2695	35	*CG9068*	973-3675	47
kl-3	1-8370	38	*CG9492*	4795-13242	54
kl-5	1-1737	44	*Dhc 93AB*	4893-6639	54
ORY	1-1989	34	*CG6059*	662-2650	54
Pp1-Y1	102-840	45	*PpN58A*	245-982	52
Pp1-Y2	1-897	43	*Pp1-87B*	230-1126	57
Ppr-Y	73-1686	34	*CG13125-PA*	97-1686	50
Ppr-Y	492-1686	35	*CG13125-PB*	90-1263	52

parts of the chromosomes. When a natural DNA sequence variant arises that substitutes C for T or G for A in the third position of a codon, any selective advantage provided by that variant is likely to be extremely small because the majority of such changes encode the same or similar amino acids. Since most mutations are expected to be deleterious, infrequent mutations providing only a slight selective advantage are likely to be lost because of the stronger selection against a number of linked mutations that are disadvantageous (Charlesworth *et al.* 1993). Alternatively, if a single highly advantageous mutation were to arise, it could be rapidly swept to fixation and carry with it all tightly linked variants, regardless of whether they were beneficial or not (Maynard Smith and Haigh 1974). Both scenarios illustrate the inability of slightly advantageous mutations to increase in frequency unless they can be separated from the effects of neighboring mutations by recombination. Considering that T-to-C or A-to-G changes in the third position of a codon have, in general, very low or no adaptive value at the protein level, the existence of codon preferences is thought to be an adaptation for more efficient translation. In support of this, highly expressed genes of Drosophila tend to contain codons ending in C or G (Shields *et al.* 1988). However, the very slight advantage conferred by more efficient translation of a single codon can only be selected if it occurs in a region of high recombination. Thus, it is thought, the higher CG content of euchromatin reflects selection for more efficient translation in these regions of high recombination.

An alternative but not exclusive hypothesis asserts that the CG enrichment of euchromatin is a more direct consequence of recombination. DNA double-strand breaks are initiators of meiotic recombination. The mechanisms that repair the double-strand breaks, generate recombinants, and repair heteroduplex mismatches have preferences that could result in CG enrichment of regions with higher levels of recombination (Brown and Jiricny 1988; Holmes *et al.* 1990; Varlet *et al.* 1990, 1996; Bill *et al.* 1998; Smith and Nicolas 1998; Nickoloff *et al.* 1999; Petranovic *et al.* 2000; Birdsell 2002).

The distribution of recombination might then have a dual influence over DNA base composition: the lack of recombination in heterochromatin impedes selection for preferred codons, leading to a lower CG content when compared with regions having normal rates of recombination. Additionally, characteristics of the recombination mechanism might itself favor an increase in CG content (Marais 2003).

The heterochromatic *TNC* genes of both *D. melanogaster* show reduced codon bias, consistent with the hypothesis discussed above. However, if selection for favored codons did not operate in regions with low recombination, we might expect to see the coding sequences of genes in these regions equilibrate in the range of 50% CG content. Such is not the case for the heterochromatic *TNC* genes of *D. melanogaster*, which have a CG content of <50%. Furthermore, for several of the heterochromatic regions, we surveyed the CG content of the most heterochromatin-proximal genes tends to be <50%. This raises the question of whether other mechanisms are responsible for CG depletion of genes in or near heterochromatin in *D. melanogaster*.

We suggest that the reduction in CG content in and near heterochromatin in *D. melanogaster* may be, at least partly, a consequence of the methylation of this DNA. The existence of DNA methylation in Drosophila was not known until recently, subsequent to the identification of a putative DNA methylase gene (Hung *et al.* 1999; Tweedie *et al.* 1999; Lyko, 2001; Kunert *et al.* 2003). Chemical analysis then revealed the existence of a small amount of cytosine methylation for a short period during early development (Gowher *et al.* 2000; Lyko *et al.* 2000). The finding of methylated DNA and genes coding for putative DNA methyltransferases and methyl-DNA-binding proteins in other invertebrate species, including within the same Drosophila genus, supports the possibility of the existence of a methylated component of the genome of *D. melanogaster* (Garcia *et al.* 2007; Schaefer and Lyko 2007). Methylated cytosine residues are known to be especially prone to mutation, undergoing spontaneous deamination to T (Shen *et al.* 1994). The elevated susceptibility of methylated cytosine to mutation can explain the CG depletion in heterochromatin if this region of the chromosome is methylated at a higher rate than the rest of the genome.

It is not yet known whether methylation occurs primarily in heterochromatin in Drosophila. We propose that it does, occurring preferentially in α-heterochromatin with a graded reduction throughout the transition into euchromatin. Several lines of indirect evidence support this proposal. First, excessive DNA methylation, produced by expression of a mouse DNA methylase in Drosophila, produces general heterochromatin-like phenotypes that are relieved by a suppressor-of-variegation mutation (Weissmann *et al.* 2003). This suggests that DNA methylation might play an important role in the formation of heterochromatin. Second, DNA methylation plays a major role in genomic imprinting in mammals (Reik and Walter 2001), and genomic imprinting appears to be specifically a phenomenon of heterochromatin in Drosophila (Maggert and Golic 2002; Ashburner *et al.* 2004). If DNA methylation also has a role in Drosophila imprinting, then it is likely that methylation occurs primarily or entirely within heterochromatin. Finally, in Arabidopsis, it has been shown that a gradient of decreasing cytosine methylation exists within the region of transition from heterochromatin to euchromatin on the left arm of chromosome 5 (Mathieu *et al.* 2002). Though these two species are quite distant, a similar methylation gradient occurring in Drosophila's

genome would correspond to, and possibly account for, the CG depletion gradient that we observed for genes in this region. In the future, the availability of more sensitive methods for detecting 5-methylcytosine should allow direct testing of this hypothesis.

Several investigators have examined trends in sequence substitutions by analyzing DNA sequences of different species of the Drosophila genus (Akashi 1996; Rodriguez-Trelles *et al.* 1999, 2000; Takano-Shimizu 1999, 2001; Bachtrog 2003; Powell *et al.* 2003; Kern and Begun 2004; Ko *et al.* 2006). The results of Takano-Shimizu (2001) and Ko *et al.* (2006) seem very interesting to us, because they show there exists a general bias in AT-increasing substitutions in the tip of the *X* chromosome of species belonging to the *D. melanogaster* species subgroup. This bias is consistent with the reduction in CG content in heterochromatic regions in general and the telomeric region of the *X* chromosome in particular. On the other hand, those studies also showed that within the *X* telomeric regions, there are restricted intervals of DNA that have strong biases of CG-increasing substitutions in the *D. teissieri–D. yakuba* and *D.erecta–D. orena* lineages. Those CG-increasing biases coincide with regional increments of recombination rates, and we already discussed the dual contribution recombination might have to elevate CG content. If our hypothesis that DNA methylation is responsible for CG reduction around heterochromatin in *D. melanogaster* is correct, it seems then that more than one mechanism might contribute to the base composition of genomes. Regional biases in the action of such mechanisms as recombination or DNA methylation may define discrete regions with different compositions.

A somewhat controversial aspect of our hypothesis is related to the role of Dnmt2. Enzymes that belong to the Dnmt2 subfamily exhibit a high degree of conservation and are the eukaryotic DNA methyltransferases with the broadest phylogenetic distribution (Goll and Bestor 2005; Ponger and Li 2005). This may be taken as evidence of the importance of these enzymes, yet elimination of Dnmt2 function does not have a large effect on viability or fertility (Pinarbasi *et al.* 1996; Okano *et al.* 1998; Kunert *et al.* 2003; Fisher *et al.* 2004; Gutierrez and Sommer 2004; Kuhlmann *et al.* 2005; Lin *et al.* 2005; Goll *et al.* 2006). However, the biggest question about Dnmt2 is related to its enzymatic function. In recent years, a number of reports have emerged purporting to demonstrate that Dnmt2 methylates DNA (Pinarbasi *et al.* 1996; Hermann *et al.* 2003; Kunert *et al.* 2003; Liu *et al.* 2003; Narsa Reddy *et al.* 2003; Tang *et al.* 2003; Fisher *et al.* 2004; Mund *et al.* 2004; Kuhlmann *et al.* 2005; Ferres-Marco *et al.* 2006; Katoh *et al.* 2006) or that it does not methylate DNA (Okano *et al.* 1998; Van Den Wyngaert *et al.* 1998; Yoder and Bestor 1998; Dong *et al.* 2001, Goll *et al.* 2006). Yet, more recently it has been shown that Dnmt2 methylates RNA, not DNA, *in vitro* (Goll *et al.* 2006, Rai *et al.* 2007) and has an *in vivo* role in zebrafish that is more consistent with RNA methylation than DNA methylation (Rai *et al.* 2007). Interestingly, Jeffery and Nakielny (2004) showed that mammal Dnmt3, an established DNA methyltransferase, was able to bind a small interfering RNA molecule *in vitro*, while mammalian Dnmt2 did not. At the moment, it is not clear what role these various interactions play in mediating the biological role of the enzymes. We consider it possible that Dnmt2 has DNA methylation activity but that it requires a cofactor that was absent from the *in vitro* preparations. Alternatively, our hypothesis that DNA methylation has played a role in the evolution of heterochromatic DNA sequence in Drosophila does not depend specifically on Dnmt2 having cytosine methylation activity. It is sufficient that there is a cytosine DNA methylation activity in *D. melanogaster*, even if the gene encoding that function has not been positively identified.

The function of DNA methylation in Drosophila: One of the functions that has been attributed to DNA methylation is the control of transposable elements (Yoder *et al.* 1997; Bird 2002; Galagan and Selker 2004; Goll and Bestor 2005; Ponger and Li 2005). In *D. melanogaster*, we propose that DNA methylation may be responsible for inducing sequence alterations in heterochromatin, where transposable elements are found in high quantity (Ashburner *et al.* 2004). Actually, several of the sequences found to be methylated in Drosophila adults are transposable elements or heterochromatic repeats (Salzberg *et al.* 2004). In *Dictyostelium discoideum*, it has been reported that Dnmt2 methylates the DNA sequence of several transposable elements (Kuhlmann *et al.* 2005; Katoh *et al.* 2006).

A similar role for DNA methylation may be found in *Neurospora crassa*, where repeated sequences are subject to mutation and silencing, a process termed RIP, for repeat-induced point mutation (for review, see Galagan and Selker 2004). A putative DNA methylase, encoded by the *rid* gene, is required for the mutagenic process and may function to generate methylated cytosines, which are then subject to a high rate of deamination leading to sequence alteration. This process has been extremely efficient; no intact transposons have been identified in the *N. crassa* genome. The sequence alterations mediated by RIP are especially prominent in the region of centromeric heterochromatin in Neurospora. This is very similar to our finding of strong CG reduction for coding sequences located in proximity to centric heterochromatin. The CG reduction exhibited by genes in this region might then be viewed as an accident of their proximity to large numbers of transposon sequences.

Transposable elements are a clear and ubiquitous danger to all genomes (Kidwell and Lisch 1997). Dnmt2 might provide an important level of protection from transposons, but its absence from many species (Ponger and Li 2005) indicates that there are mechanisms other than mutagenesis that organisms use to

combat transposons, including transcriptional and post-transcriptional silencing (Bird 2002; Goll and Bestor 2005; Cerutti and Casas-Molano 2006). To understand the degree to which these mechanisms may act to control transposons and the extent of their cooperation in this process is likely to be an interesting area of future investigation.

In this article, we showed that paralogous comparisons are a powerful tool for understanding the way genes and genomes evolve. The study of a single gene family formed by just five genes in *D. melanogaster* permitted us to detect possible signatures of the activity of a particular sequence-altering mechanism acting in specific chromosome regions. We realize that our conclusions remain speculative and need confirmation. In the coming years, many more species of the Drosophila genus will have their genomes sequenced. We have some data about the *TNC* gene family in 11 more Drosophila species, including *D. pseudoobscura*, whose genome was already published (Richards *et al.* 2005). Unfortunately, at this time, the quality of the assembly of the genome sequences in these species does not allow us to be certain of the locations of all the *TNC* genes in these species, which is why analyses of these genes are not included in this article.

Another kind of analysis that will help us to validate our hypothesis will be to identify and examine other gene families with members located in heterochromatin and euchromatin or even single genes that have changed location from euchromatin to heterochromatin. For these purposes, having accurately assembled genome sequences of additional Drosophila species will be extremely valuable.

We thank Roberto Marco and Alfredo Villasante for their help in the first stages of the study here presented. This work was supported by NIH grant GM065604.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

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.

Anderson, E. G., 1925 Crossing over in a case of attached *X* chromosomes in *Drosophila melanogaster*. Genetics **10:** 403–417.

Ashburner, M., K. G. Golic and R. S. Hawley, 2004 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bachtrog, D., 2003 Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila Miranda*. Genetics **165:** 1221–1232.

Baker, W. K., and A. Rein, 1962 The dichotomous action of *Y* chromosomes on the expression of position-effect variegation. Genetics **47:** 1399–1407.

Bernardi, G., 1989 The isochore organization of the human genome. Annu. Rev. Genet. **23:** 637–661.

Bernardi, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

Biggs, W. H. III, K. H. Zavitz, B. Dickson, A. van der Straten, D. Brunner *et al.*, 1994 The Drosophila *rolled* locus encodes a MAP kinase required in the sevenless signal transduction pathway. EMBO J. **13:** 1628–1635.

Bill, C. A., W. A. Duran, N. R. Miselis and J. A. Nickoloff, 1998 Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics **149:** 1935–1943.

Bird, A., 2002 DNA methylation patterns and epigenetic memory. Genes Dev. **16:** 6–21.

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

Bradnam, K. R., C. Seoighe, P. M. Sharp and K. H. Wolfe, 1999 G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. Mol. Biol. Evol. **16:** 666–675.

Brown, T. C., and J. Jiricny, 1988 Different base/base mispairs are correlated with different efficiencies and specificities in monkey kidney cells. Cell **54:** 705–711.

Carulli, J. P., D. E. Krane, D. L. Hartl and H. Ochman, 1993 Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics **134:** 837–845.

Carvalho, A. B., B. A. Dobo, M. D. Vibranovski and A. G. Clark, 2001 Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **98:** 13225–13230.

Cerutti, H., and J. A. Casas-Mollano, 2006 On the origin of RNA-mediated silencing: from protist to man. Curr. Genet. **50:** 81–99.

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

Colot, V., and J. L. Rossignol, 1999 Eukaryotic DNA methylation as an evolutionary device. BioEssays **21:** 402–411.

Cooper, D. N., and H. Yousouffian, 1988 The CpG dinucleotide and human genetic disease. Hum. Genet. **78:** 151–155.

Coulondre, C., J. H. Miller, P. J. Farabaugh and W. Gilbert, 1978 Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274:** 775–780.

Daubin, V., and G. Perrière, 2003 G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. **20:** 471–483.

Deschavanne, P., and J. Filipski, 1995 Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E.coli* genes. Nucleic Acids Res. **23:** 1350–1353.

Devlin, R. H., B. Bingham and B. T. Wakimoto, 1990 The organization and expression of the *light* gene, a heterochromatic gene of *Drosophila melanogaster*. Genetics **125:** 129–140.

Dimitri, P., N. Corradini, F. Rossi, F. Verni, G. Cenci *et al.*, 2003 Vital genes in the heterochromatin of chromosomes *2* and *3* of *Drosophila melanogaster*. Genetica **117:** 209–215.

Dong, A., J. A. Yoder, X. Zhang, L. Zhou, T. H. Bestor *et al.*, 2001 Structure of human DNMT2, and enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. Nucleic Acids Res. **29:** 439–448.

Dujon, B., D. Alexandraki, B. Andre, W. Ansorge, V. Baladron *et al.*, 1994 Complete DNA sequence of yeast chromosome *XI*. Nature **369:** 371–378.

Eberl, D. F., B. J. Duyff and A. J. Hilliker, 1993 The role of heterochromatin in the expression of a heterochromatic gene, the *rolled* locus of *Drosophila melanogaster*. Genetics **134:** 277–292.

Eyre-Walker, A., and L. D. Hurst, 2001 The evolution of isochores. Nat Rev Genet **2:** 549–555.

Feldmann, H., M. Aigle, G. Aljinovic, B. Andre, M. C. Baclet *et al.*, 1994 Complete DNA sequence of yeast chromosome *II*. EMBO J. **13:** 5795–5809.

Ferres-Marco, D., I. Gutierrez-Garcia, D. M. Vallejo, J. Bolivar, F. J. Gutierrez-Aviño *et al.*, 2006 Epigenetic silencers and Notch collaborate to promote malignant tumours by *Rb* silencing. Nature **439:** 430–436.

Filatov, V. L., A. G. Katrukha, T. V. Bulargina and N. B. Gusev, 1999 Troponin: structure, properties, and mechanism of functioning. Biochemistry **64:** 969–985.

Fisher, O., R. Siman-Tov and S. Ankri, 2004 Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase (Ehmeth) in the protozoan parasite *Entamoeba histolytica*. Nucleic Acids Res. **32:** 287–297.

FlyBase, 1999 The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res. **27:** 85–88.

Fryxell, K. J., and E. Zuckerkandl, 2000 Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. **17:** 1371–1383.

Fullerton, S. M., A. Bernardo Carvalho and A. G. Clark, 2001 Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. **18:** 1139–1142.

Galagan, J. E., and E. U. Selker, 2004 RIP: the evolutionary cost of genome defense. Trends Genet. **20:** 417–423.

Garcia, R. N., M. F. D'ávila, L. J. Robe, E. L. Loreto, Y. Panzera *et al.*, 2007 First evidence of methylation in the genome of *Drosophila willistoni*. Genetica **131:** 91–105.

Gatti, M., and S. Pimpinelli, 1992 Functional elements in *Drosophila melanogaster* heterochromatin. Annu. Rev. Genet. **26:** 239–275.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown *et al.*, 2000 Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **97:** 11383–11390.

Goll, M. G., and T. H. Bestor, 2005 Eukaryotic cytosine DNA methyltransferases. Annu. Rev. Biochem. **74:** 481–514.

Goll, M. G., F. Kirpekar, K. A. Maggert, J. A. Yoder, C. -L. Hsieh *et al.*, 2006 Methylation of $tRNA^{Asp}$ by the DNA methyltransferase homolog Dnmt2. Science **311:** 395–398.

Gowher, H., O. Leismann and A. Jeltsch, 2000 DNA of *Drosophila melanogaster* contains 5-methylcytosine. EMBO J. **19:** 6918–6923.

Graur, D., and W. -H. Li, 2000 *Fundamentals of Molecular Evolution*, Ed. 2. Sinauer Associates, Sunderland, MA.

Greenblatt, M. S., W. P. Bennett, M. Hollstein and C. C. Harris, 1994 Mutations in the *p53* tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Res. **54:** 4855–4878.

Gutierrez, A., and R. J. Sommer, 2004 Evolution of *dnmt*-2 and *mbd*-2-like genes in the free-living nematodes *Pristionchus pacificus*, *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res. **32:** 6388–6396.

Hearn, M. G., A. Hedrick, T. A. Grigliatti and B. T. Wakimoto, 1991 The effect of modifiers of position-effect variegation on the variegation of heterochromatic genes of *Drosophila melanogaster*. Genetics **128:** 785–797.

Heitz, E., 1934 Uber- und β-Heterochromatin sowie Konstanz und Bau der Chromomeren bei Drosophila. Biol. Zentbl. **54:** 588–609.

Henikoff, S., K. Ahmad and H. S. Malik, 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. Science **293:** 1098–1102.

Hermann, A., S. Schmitt and A. Jeltsch, 2003 The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J. Biol. Chem. **278:** 31717–31721.

Herranz, R., C. Díaz-Castillo, T. P. Nguyen, T. L. Lovato, R. M. Cripps *et al.*, 2004 Expression patterns of the whole *troponin C* gene repertoire during Drosophila development. Gene Expr. Patterns **4:** 183–190.

Herranz, R., J. Mateos and R. Marco, 2005 Diversification and independent evolution of troponin C genes in insects. J. Mol. Evol. **60:** 31–44.

Hessler, A. Y., 1958 V-type position effects at the *light* locus in *Drosophila melanogaster*. Genetics **43:** 395–403.

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

Holmes, J. Jr., S. Clark and P. Modrich, 1990 Strand-specific mismatch correction in nuclear extracts of human and *Drosophila melanogaster* cell lines. Proc. Natl. Acad. Sci. USA **87:** 5837–5841.

Hung, M. S., N. Karthikeyan, B. Huang, H. C. Koo, J. Kiger *et al.*, 1999 Drosophila proteins related to vertebrate DNA (5-cytosine) methyltransferases. Proc. Natl. Acad. Sci. USA **96:** 11940–11945.

Ikemura, T., and K. Wada, 1991 Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res. **19:** 4333–4339.

Jeffery, L., and S. Nakielny, 2004 Components of the DNA methylation system of chromatin control are RNA-binding proteins. J. Biol. Chem. **279:** 49479–49487.

Jones, P.A., J. D. Buckley, B. E. Henderson, R. K. Ross and M. C. Pike, 1991 From gene to carcinogen: a rapidly evolving field in molecular epidemiology. Cancer Res. **51:** 3617–3620.

Katoh, M., T. Curk, Q. Xu, B. Zupan, A. Kuspa *et al.*, 2006 Developmentally regulated DNA methylation in *Dictyostelium discoideum*. Eukaryotic Cell **5:** 18–25.

Kern, A. D., and D. J. Begun, 2004 Patterns of polymorphism and divergence from noncoding sequences of *Drosophila melanogaster* and *D. simulans*: evidence for nonequilibrium processes. Mol. Biol. Evol. **22:** 51–62.

Khvostova, V. V., 1939 The role played by the inert chromosome regions in the position effect of the *cubitus interruptus* gene in *Drosophila melanogaster*. Izv. Akad. Nauk. SSSR **1939:** 541–574.

Kidwell, M. G., and D. Lisch, 1997 Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. USA **94:** 7704–7711.

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

Ko, W. -Y., S. Piao and H. Akashi, 2006 Strong region-specific heterogeneity in base composition evolution on the Drosophila *X* chromosome. Genetics **174:** 349–362.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Kuhlmann, M., B. E. Borisova, M. Kaller, P. Larsson, D. Stach *et al.*, 2005 Silencing of retrotransposons in *Dictyostelium* by DNA methylation and RNAi. Nucleic Acids Res. **33:** 6405–6417.

Kunert, N., J. Marhold, J. Stanke, D. Stach and F. Lyko, 2003 A Dnmt2-like protein mediates DNA methylation in Drosophila. Development **130:** 5083–5090.

Lin, M. -J., L. -Y. Tang, M. Narsa Reddy and C. -K. Shen, 2005 DNA methyltransferase gene *dDnmt2* and longevity of *Drosophila*. J. Biol. Chem. **280:** 861–864.

Liu, K., Y. F. Wang, C. Cantemir and M. T. Muller, 2003 Endogenous assays of DNA methyltransferases: evidence for differential activities of DNMT1, DNMT2, and DNMT3 in mammalian cells in vitro. Mol. Cell. Biol. **23:** 2709–2719.

Lu, B. Y., P. C. Emtage, B. J. Duyf, A. J. Hilliker and J. C. Eissenberg, 2000 Heterochromatin protein 1 is required for the normal expression of two heterochromatin genes in Drosophila. Genetics **155:** 699–708.

Lyko, F., 2001 DNA methylation learns to fly. Trends Genet. **17:** 169–172.

Lyko, F., B. H. Ramsahoye and R. Jaenisch, 2000 DNA methylation in *Drosophila melanogaster*. Nature **408:** 538–540.

Maggert, K. A., and K. G. Golic, 2002 The *Y* chromosome of *Drosophila melanogaster* exhibits chromosome-wide imprinting. Genetics **162:** 1245–1258.

Marais, G., 2003 Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19:** 330–338.

Marais, G., and G. Piganeau, 2002 Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol. Biol. Evol. **19:** 1399–1406.

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

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

Mathieu, O., G. Picard and S. Tourmente, 2002 Methylation of a heterochromatin-euchromatin transition region in *Arabidopsis thaliana* chromosome *5* left arm. Chromosome Res. **10:** 455–466.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **231:** 1114–1116.

Meunier, J., and L. Duret, 2004 Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. **21:** 984–990.

Muller, H. J., and T. S. Painter, 1932 The differentiation of the sex chromosomes of *Drosophila* into genetically active and inert regions. Z. Indukt. Abstammungs Vererbungsl. **62:** 316–365.

Mund, C., T. Musch, M. Strodicke, B. Assmann, E. Li *et al.*, 2004 Comparative analysis of DNA methylation patterns in transgenic *Drosophila* overexpressing mouse DNA methyltransferases. Biochem. J. **378:** 763–768.

Narsa Reddy, M., L. -Y. Tang, T. -L. Lee and C. -K. Shen, 2003 A candidate gene for *Drosophila* genome methylation. Oncogene **22:** 6301–6303.

Nickoloff, J. A., D. B. Sweetser, J. A. Clikeman, G. J. Khalsa and S. L. Wheeler, 1999 Multiple heterologies increase mitotic double-strand break-induced allelic gene conversion tract lengths in yeast. Genetics **153:** 665–679.

Okano, M., S. Xie and E. Li, 1998 Dnmt2 is not required for *de novo* and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Res. **26:** 2536–2540.

Petranovic, M., K. Vlahovic, D. Zahradka, S. Dzidc and M. Radman, 2000 Mismatch repair in xenopus egg extracts is not strand-directed by DNA methylation. Neoplasma **47:** 375–381.

Pinarbasi, E., J. Elliott and D. Hornby, 1996 Activation of a pseudo DNA methyltransferase by deletion of a single amino acid. J. Mol. Biol. **257:** 804–813.

Ponger, L., and W. -H. Li, 2005 Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. Mol. Biol. Evol. **22:** 1119–1128.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, New York.

Powell, J. R., E. Sezzi, E. Moriyama, J. Gleason, and A. Caccone, 2003 Analysis of shift in codon usage in Drosophila. J. Mol. Evol. **57:** S214–225.

Rai, K., S. Chidester, C. V. Zavala, E. J. Manos, S. R. James *et al.*, 2007 Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. Genes Dev. **21:** 261–266.

Reik, W., and J. Walter, 2001 Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. **2:** 21–32.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. **15:** 1–18.

Rodriguez-Trelles, F., R. Tarrio and F. J. Ayala, 1999 Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics **153:** 339–350.

Rodriguez-Trelles, F., R. Tarrio and F. J. Ayala, 2000 Fluctuating mutation bias and the evolution of base composition in *Drosophila*. J. Mol. Evol. **50:** 1–10.

Salzberg, A., O. Fisher, R. Siman-Tov and S. Ankri, 2004 Identification of methylated sequences in genomic DNA of adult *Drosophila melanogaster*. Biochem. Biophys. Res. Commun. **322:** 465–469.

Schaefer, M., and F. Lyko, 2007 DNA methylation with a sting: an active DNA methylation system in the honeybee. BioEssays **29:** 208–211.

Schultz, J., 1936 Variegation in *Drosophila* and the inert chromosome regions. Proc. Natl. Acad. Sci. USA **22:** 27–33.

Selker, E. U., 1990 Premeiotic instability of repeated sequences in *Neurospora crassa*. Annu. Rev. Genet. **24:** 579–613.

Sharp, P. M., and A. T. Lloyd, 1993 Regional base composition variation along yeast chromosome *III*: evolution of chromosome primary structure. Nucleic Acids Res. **21:** 179–183.

Sharp, P. M., D. C. Shields, K. H. Wolfe and W. H. Li, 1989 Chromosomal location and evolutionary rate variation in enterobacterial genes. Science **246:** 808–810.

Shen, J. C., W. M. Rideout, III and P. A. Jones, 1994 The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. **22:** 972–976.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Singer, M. J., B. A. Marcotte and E. U. Selker, 1995 DNA methylation associated with repeat-induced point mutation in *Neurospora crassa*. Mol. Cell Biol. **15:** 5586–5597.

Smith, K. N., and A. Nicolas, 1998 Recombination at work for meiosis. Curr. Opin. Genet. Dev. **8:** 200–211.

Sun, F. L., M. H. Cuaycong, C. A. Craig, L. L. Wallrath, J. Locke *et al.*, 2000 The fourth chromosome of *Drosophila melanogaster*: interspersed euchromatic and heterochromatic domains. Proc. Natl. Acad. Sci. USA **97:** 5340–5345.

Sun, F. L, K. Haynes, C. L. Simpson, S. D. Lee, L. Collins *et al.*, 2004 *cis*-Acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. Mol. Cell Biol. **24:** 8210–8220.

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

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

Tang, L. Y., M. Narsa Reddy, V. Rasheva, T. L. Lee, M. -J. Lin *et al.*, 2003 The eukaryotic DNMT2 genes encode a new class of cytosine-5 DNA methyltransferases. J. Biol. Chem. **278:** 33613–33616.

Tulin, A., D. Stewart and A. C. Spradling, 2002 The Drosophila heterochromatic gene encoding poly(ADP-ribose) polymerase (PARP) is required to modulate chromatin structure during development. Genes Dev. **16:** 2108–2119.

Tweedie, S., H. H. Ng, A. L. Barlow, B. M. Turner, B. Hendrich *et al.*, 1999 Vestiges of a DNA methylation system in *Drosophila melanogaster?* Nat. Genet. **23:** 389–390.

Van den Wyngaert, I., J. Sprengel, S. U. Kass and W. H. M. L. Luyten, 1998 Cloning and analysis of a novel human putative DNA methyltransferase. FEBS Lett **426:** 283–289.

Varlet, I., M. Radmana and P. Brooks, 1990 DNA mismatch repair in Xenopus egg extracts: repair efficiency and DNA repair synthesis for all single base-pair mismatches. Proc. Natl. Acad. Sci. USA **87:** 7883–7887.

Varlet, I., B. Canard, P. Brooks, G. Cerovic and M. Radman, 1996 Mismatch repair in Xenopus egg extracts: DNA strand breaks act as signals rather than excision points. Proc. Natl. Acad. Sci. USA **93:** 10156–10161.

Wakimoto, B. T., and M. G. Hearn, 1990 The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome *2L* of *Drosophila melanogaster*. Genetics **125:** 141–154.

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

Watters, M. K., T. A. Randall, B. S. Margolin, E. U. Selker and D. R. Stadler, 1999 Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in Neurospora. Genetics **153:** 705–714.

Weiler, K. S., and B. T. Wakimoto, 2002 Suppression of heterochromatic gene variegation can be used to distinguish and characterize *E(var)* genes potentially important for chromosome structure in *Drosophila melanogaster*. Mol. Genet. Genomics **266:** 922–932.

Weissmann, F., I. Muyrers-Chen, T. Musch, D. Stach, M. Wiessler *et al.*, 2003 DNA hypermethylation in *Drosophila melanogaster* causes irregular chromosome condensation and dysregulation of epigenetic histone modifications. Mol. Cell Biol. **23:** 2577–2586.

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

Yang, A. S., P. A. Jones and A. Shibata, 1996 The mutational burden of 5-methylcytosine, pp. 77–94 in *Epigenetic Mechanisms of Gene Regulation*, edited by V. E. A. Russo, R. A. Martienssen and A. D. Riggs. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Yoder, J. A., and T. H. Bestor, 1998 A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. Hum. Mol. Genet. **7:** 279–284.

Yoder, J. A., C. P. Walsh and T. H. Bestor, 1997 Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. **13:** 335–340.

Zhang, J., A. M. Dean, F. Brunet and M. Long, 2004 Evolving protein functional diversity in new genes of *Drosophila*. Proc. Natl. Acad. Sci. USA **101:** 16246–16250.

Zhang, R., and C. -T. Zhang, 2004 Isochore structures in the genome of the plant *Arabidopsis thaliana*. J. Mol. Evol. **59:** 227–238.

Communicating editor: G. Gibson