

## Patterns of Base Composition Within the Genes of *Drosophila melanogaster*

Richard M. Kliman,<sup>1</sup> Adam Eyre-Walker<sup>2</sup>

<sup>1</sup>Department of Biology, Radford University, Radford, VA 24142, USA

<sup>2</sup>Centre for the Study of Evolution and Department of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK

Received: 24 February 1997 / Accepted: 10 November 1997

**Abstract.** Base composition is not uniform across the genome of *Drosophila melanogaster*. Earlier analyses have suggested that there is variation in composition in *D. melanogaster* on both a large scale and a much smaller, within-gene, scale. Here we present analyses on 117 genes which have reliable intron/exon boundaries and no known alternative splicing. We detect significant heterogeneity in G+C content among intron segments from the same gene, as well as a significant positive correlation between the intron and the third codon position G+C content within genes. Both of these observations appear to be due, in part, to an overall decline in intron and third codon position G+C content along *Drosophila* genes with introns. However, there is also evidence of an increase in third codon position G+C content at the start of genes; this is particularly evident in genes without introns. This is consistent with selection acting against preferred codons at the start of genes.

**Key words:** G+C content — Isochores — Synonymous codon use

### Introduction

Given that genomes contain information, it is not surprising that base composition is not uniform along chromosomes. However, the scale over which compositional variation is observed, and the fact that it is detected in

DNA thought largely to be noncoding, has generated considerable interest. Most attention has focused on large-scale variation along chromosomes, thought until recently to be restricted to mammals. With the sequencing of yeast chromosomes (Oliver et al. 1992; Dujon et al. 1994; Feldmann et al. 1994) and the rapid accumulation of other DNA sequence data, it now appears that variation in base composition along chromosomes is widespread; it has been documented in mammals (Bernardi et al. 1985; Bernardi 1989), *Drosophila* (Carulli et al. 1993; Kliman and Hey 1994), yeast (Sharp and Lloyd 1992; Dujon et al. 1994; Feldmann et al. 1994), and bacteria (Deschavanne and Filipinski 1995).

While genomewide variation in base composition is common, the scales of variation differ among organisms. In mammals, G+C content varies over megabases of DNA (Gardiner et al. 1990; Pilia et al. 1993), whereas in *Saccharomyces cerevisiae*, the variation is over a scale of about 50 to 100 kilobases (kb) (Sharp and Lloyd 1992; Dujon et al. 1994; Feldmann et al. 1994), similar to that found in *Escherichia coli* (Deschavanne and Filipinski 1995). The scale in *Drosophila melanogaster* remains unclear. Carulli et al. (1993) found substantial variation in G+C content among 140- to 340-kb fragments of *Drosophila* chromosomes, suggesting that ~100 kb was the lower limit for the variation.

Compositional heterogeneity has also been observed at a much smaller scale, though examples are rare. In mammals there are “CpG islands” (Bird 1987). These are kilobase stretches of DNA with relatively high levels of CpG dinucleotides and G+C content. The origin and function of these sequences remain obscure, though their lack of methylation is implicated in gene control (Lewis

and Bird 1991). In *E. coli*, there is a high frequency of A and a lower frequency of G at both ends of genes (Eyre-Walker and Bulmer 1993; Eyre-Walker 1996). This is thought to be due to constraints imposed by ribosome binding.

Compositional heterogeneity has also been detected within the genes of *D. melanogaster*. In a sample of 79 genes with multiple introns, 33 showed significant heterogeneity among introns in G+C content (Kliman and Hey 1994). Here we build on the earlier analyses using a new data set and consider three aspects of compositional heterogeneity. We establish that there is significant compositional heterogeneity in intron composition within genes. We then investigate (i) whether there is a correlation between intron and exon G+C content within genes and (ii) whether there are trends in base composition along genes.

## Materials and Methods

### Selection of Genes

In *D. melanogaster*, like many other organisms, exons have a substantially higher G+C content than introns. If boundaries are assigned incorrectly, then artifactual heterogeneity and spurious correlations of intron and exon base composition can arise. There are also potential complications with genes that are alternatively spliced. Therefore, we restricted our analyses to genes that met the following criteria: (1) there were no published reports of alternative splicing for introns and exons between the start and the stop codon; (ii) the entire genomic DNA sequence was determined between the start and the stop codons; and (iii) intron/exon boundaries were determined experimentally, usually by comparison of genomic and cDNA sequences. Of 117 suitable genes, 47 had one intron and 39 had at least two introns (see Table 1). We did not remove genes that are members of multigene families (e.g., actins) since they appear to be the products of fairly ancient duplications. They are, therefore, independent with respect to codon usage and intron G+C content.

### Statistical Analysis—Base Composition

We divided each intron in half, allowing us to include genes with one intron. To test for heterogeneity in intron G+C content within genes, we performed  $G$  tests of independence (with William's correction) to test the null hypothesis of G+C content equanimity within genes. Like chi-square values,  $G$  values can be summed across tests, along with degrees of freedom, to yield an overall test of heterogeneity.

The intron segments were paired with half of the adjacent exon (except that terminal exons were left intact) to form an array of intron/exon pairs. Analyses were performed only on amino acid-coding regions of exons and ignored the universally conserved start codon. To test if there was a correlation between paired intron and exon G+C content within genes, we calculated a statistic related to Pearson's correlation coefficient,

$$r_{\text{within}} = \frac{\sum_i^G \sum_j^R (E_{ij} - \bar{E}_i)(I_{ij} - \bar{I}_i)}{\sqrt{\sum_i^G \sum_j^R (E_{ij} - \bar{E}_i)^2 \sum_i^G \sum_j^R (I_{ij} - \bar{I}_i)^2}} \quad (1)$$

where  $E_{ij}$  and  $I_{ij}$  are the exon (first, second, or third codon position) and intron G+C contents of the  $j$ th of  $R$  segments of the  $i$ th of  $G$  genes. By subtracting the gene mean of the intron and exon G+C contents from each point, the statistic removes any correlation between intron and exon G+C content that may exist among genes. The statistic can be thought of as the correlation coefficient for a series of parallel lines, one for each gene, running through the data; its square, like that of the standard correlation coefficient, is the proportion of variance within genes explained by the parallel lines. While the significance of  $r_{\text{within}}$  can be assessed by conventional statistical techniques, to avoid problems caused by departures from the assumptions of parametric methods, we employed a Monte Carlo approach. We reshuffled the  $x$  (intron) and  $y$  (exon) values within genes to yield a randomized data set of points [e.g.,  $((x_{11}, y_{15}), (x_{12}, y_{11}), \dots), ((x_{24}, y_{21}), (x_{21}, y_{27}), \dots), \dots$ , etc.], then recalculated  $r_{\text{within}}$ . Randomization was performed 100,000 times per analysis. Since we are testing for a positive correlation, we calculated the proportion of times that  $r_{\text{within}}$  for the randomized data exceeded  $r_{\text{within}}$  for the original data; i.e., we performed one-tailed tests.

To test for trends in base composition and codon usage along genes, we performed a linear regression analysis of base composition against distance from the start codon, removing base composition differences among genes by subtracting gene means (as above). To maximize the amount of information available for an analysis, we considered each nucleotide site individually; i.e., G+C content (or preferred codon usage) was either 0 or 1. In these analyses we also looked at preferred codon use, where preferred codons are defined as those that increase in frequency as codon bias in *D. melanogaster* genes increases (Akashi 1995). The distance from the start codon was calculated as base pairs. We calculated the test statistic,

$$b_{\text{within}} = \frac{\sum_i^G \sum_j^R (K_{ij} - \bar{K}_i)(D_{ij} - \bar{D}_i)}{\sqrt{\sum_i^G \sum_j^R (D_{ij} - \bar{D}_i)^2}} \quad (2)$$

where  $D_{ij}$  is the distance of the  $j$ th point from the start codon of the  $i$ th gene, and  $K$  is the G+C content (for either intron, first, second, or third codon position) or preferred codon usage. The test statistic is the slope of the least-squares line through the data after subtracting the (gene) mean G+C content and distance from each point. It is also the slope of a series of parallel lines, one for each gene, drawn through the data.

Unfortunately, this statistic is disproportionately influenced by long genes (with weighting approximately proportional to the gene length cubed). To overcome this, we repeated the analysis after scaling the distance along the gene so that it ran from 0, the first base of the start codon, to 1, the last base of the stop codon. This weights genes approximately proportional to their length. The significance of slopes was assessed by Monte Carlo methods analogous to those used in the correlation analysis (performing 1,000 randomizations). Probabilities were estimated as two-tailed values.

### Statistical Analysis—Substitution Rates

We also examined the rate of synonymous substitution along *Drosophila* genes using a data set of 33 aligned genes from *D. melanogaster* and *D. pseudoobscura*. These were kindly provided by Dr. Et-suko Moriyama. The analysis was restricted to codons where the amino acid was the same in the two species. The average synonymous site divergence was calculated for each codon along the genes as

$$K_j = \frac{\sum_i^{n_j} (Q_{ij} - \bar{Q}_i)}{n_j} \quad (3)$$

**Table 1.** Genes used

Locus	GenBank accession No.	Codon adaptation index	Codons	Introns	Intron bases
<i>Act42A</i>	K00670	0.471	376	0	·
<i>Act87E</i>	X12452	0.671	376	0	·
<i>Ama</i>	M23561	0.481	333	0	·
<i>Amy-d</i>	X04569	0.725	494	0	·
<i>Dpt</i>	Z11728	0.38	106	0	·
<i>Efla10</i>	X06869	0.693	463	0	·
<i>HLHm5</i>	X16552	0.521	178	0	·
<i>HLHm7</i>	X16553	0.514	179	0	·
<i>Hsp67Bc</i>	X06542	0.357	169	0	·
<i>Hsp83</i>	X03810	0.715	717	0	·
<i>Mst84Db</i>	X67703	0.262	74	0	·
<i>Mst84Dc</i>	X67703	0.27	55	0	·
<i>Mst84Dd</i>	X67703	0.236	68	0	·
<i>Mst87F</i>	Y00831	0.415	56	0	·
<i>Pp1-13C</i>	X69974	0.442	302	0	·
<i>Pp1-87B</i>	X55198	0.514	302	0	·
<i>RpA1</i>	X05016	0.677	113	0	·
<i>Ser99Da</i>	M24379	0.697	265	0	·
<i>Tgfb-60</i>	M77012	0.58	455	0	·
<i>Vm26Ab</i>	M18280	0.706	141	0	·
<i>Vm32Ec</i>	M27647	0.371	116	0	·
<i>ac</i>	M17120	0.265	201	0	·
<i>aTub84D</i>	M14645	0.663	450	0	·
<i>esg</i>	M83207	0.401	470	0	·
<i>fkh</i>	J03177	0.409	510	0	·
<i>l(1)sc</i>	X12549	0.365	257	0	·
<i>m4</i>	X16551	0.675	152	0	·
<i>nullo</i>	X65444	0.433	213	0	·
<i>sc</i>	M17119	0.264	345	0	·
<i>sisA</i>	L22755	0.376	189	0	·
<i>slp1</i>	X66095	0.481	322	0	·
<i>Acp95EF</i>	M32022	0.177	52	1	58
<i>Act79B</i>	M18829	0.668	376	1	356
<i>Act88F</i>	M18830	0.681	376	1	56
<i>Anr</i>	X56726	0.271	57	1	58
<i>Bj1</i>	X58530	0.413	547	1	55
<i>CecC</i>	Z11167	0.419	63	1	65
<i>Cp15</i>	X02497	0.574	115	1	67
<i>Cp18</i>	X02497	0.609	172	1	172
<i>Cp19</i>	X02497	0.63	173	1	85
<i>Cp36</i>	X05245	0.559	286	1	87
<i>Cp38</i>	X05245	0.533	306	1	222
<i>Cyt-b5</i>	X15008	0.361	414	1	324
<i>Dhod</i>	L00964	0.358	389	1	50
<i>Dox-A2</i>	M63010	0.439	494	1	57
<i>Edg78E</i>	M71247	0.597	122	1	70
<i>Edg91</i>	M71250	0.392	159	1	59
<i>Est-6</i>	M33780	0.296	544	1	47
<i>Fbp1</i>	X69965	0.481	1,030	1	55
<i>Fbf2</i>	S57693	0.544	256	1	56
<i>Fcp3C</i>	M18281	0.294	210	1	69
<i>ImpL1</i>	M97259	0.398	341	1	70
<i>ImpL2</i>	L23066	0.553	263	1	71
<i>Lcp3</i>	V00203	0.609	112	1	52
<i>Lcp4</i>	V00203	0.595	111	1	53
<i>M(3)99D</i>	X00848	0.511	133	1	55
<i>MtnA</i>	X03758	0.653	40	1	261
<i>Sod</i>	X17332	0.592	153	1	716
<i>Tpi</i>	X57576	0.776	247	1	52

**Table 1.** Continued

Locus	GenBank accession No.	Codon adaptation index	Codons	Introns	Intron bases
<i>Uro</i>	X51940	0.441	352	1	65
<i>Yp1</i>	V00248	0.669	439	1	71
<i>aTub67C</i>	M14646	0.475	462	1	484
<i>aTub84B</i>	M14643	0.712	450	1	487
<i>amd</i>	X04695	0.418	465	1	479
<i>ems</i>	X66270	0.45	497	1	287
<i>eve</i>	M14767	0.531	376	1	67
<i>fs(1)K10</i>	X12836	0.382	463	1	850
<i>ftz</i>	X00854	0.504	413	1	146
<i>gt</i>	X61148	0.427	448	1	71
<i>kni</i>	X13331	0.472	429	1	210
<i>mex1</i>	M63626	0.497	83	1	212
<i>mus209</i>	M33950	0.545	260	1	56
<i>ninA</i>	M22851	0.519	237	1	66
<i>pn</i>	Z12141	0.311	404	1	62
<i>prd</i>	M14548	0.382	613	1	352
<i>tsl</i>	Z30342	0.379	356	1	344
<i>twi</i>	X12506	0.538	490	1	116
<i>y</i>	X04427	0.256	541	1	2,715
<i>Arf72A</i>	M61127	0.498	180	2	356
<i>Arf79F</i>	S62079	0.299	182	2	283
<i>Bsg25D</i>	X04896	0.379	741	2	1,936
<i>Bx42</i>	X64536	0.454	547	2	327
<i>Dbp73D</i>	M74824	0.269	572	2	106
<i>Efla48</i>	X06870	0.534	462	2	526
<i>Eh</i>	X72303	0.369	97	2	269
<i>M(3)67C</i>	M22142	0.707	131	2	408
<i>Mp20</i>	Y00795	0.757	184	2	199
<i>Pgd</i>	M80598	0.57	481	2	1,486
<i>Pros35</i>	X62285	0.417	279	2	118
<i>RpI135</i>	X17298	0.288	1,129	2	113
<i>RpII140</i>	X05709	0.381	1,123	2	126
<i>Sgs5</i>	X04269	0.298	163	2	107
<i>Yp3</i>	X04754	0.662	420	2	126
<i>aTub85E</i>	M14644	0.555	449	2	462
<i>bam</i>	X56202	0.343	442	2	118
<i>nos</i>	M72421	0.353	401	2	613
<i>swa</i>	X56023	0.388	548	2	261
<i>wdn</i>	M23391	0.35	868	2	162
<i>z</i>	X06743	0.405	574	2	175
<i>Arr1</i>	M30140	0.541	364	3	700
<i>GTP-bp</i>	X71866	0.495	368	3	239
<i>His2AvD</i>	X15549	0.427	141	3	1,026
<i>Rh2</i>	M12896	0.371	381	3	220
<i>RpII215</i>	M27431	0.396	1,896	3	779
<i>Top2</i>	X61209	0.4	1,447	3	192
<i>Uch</i>	X69679	0.448	227	3	266
<i>osk</i>	M63492	0.335	606	3	389
<i>ph1</i>	X07181	0.332	666	3	185
<i>ry</i>	Y00308	0.384	1,335	3	1,149
<i>su(s)</i>	M57889	0.338	1,322	4	901
<i>Dbp45A</i>	L13612	0.318	527	5	276
<i>yema</i>	X63503	0.371	1,002	5	291
<i>tld</i>	U04239	0.349	1,057	6	396
<i>Pxd</i>	X68131	0.358	690	7	395
<i>rdgC</i>	M89628	0.337	661	11	1,621
<i>tor</i>	X15150	0.337	923	13	820
<i>LanA</i>	M96388	0.388	3,712	14	2,447

where

$$\bar{Q}_i = \frac{\sum_j Q_{ij}}{l_i}$$

where  $Q_{ij}$  is 0 if the codons are identical and 1 if they differ by a synonymous difference,  $l_i$  is the length of the  $i$ th gene in codons, and  $n_j$  is the number of genes at the  $j$ th codon in the analysis. This formula calculates the average (uncorrected) synonymous divergence at each codon subtracting the gene average. We subtracted the gene average to avoid any spurious correlations caused by correlations between synonymous divergence and gene length between genes; the longest genes in our sample have some of the highest synonymous divergences.

## Results

### Intron G+C Content Within Genes

Overall, there is highly significant heterogeneity in base composition among intron segments within genes ( $G = 605.0$ ,  $df = 282$ ,  $P < 10^{-13}$ ). Individually, 18 of the 86 genes show significant heterogeneity at the 5% level. If we remove from the analysis genes with only one intron, then 13 of the 39 remaining genes show significant evidence of heterogeneity. Overall, however, the one-intron genes show significant heterogeneity ( $G = 68.5$ ,  $df = 47$ ,  $P = 0.022$ ).

The presence of within-gene variation in intron G+C content raises two immediate questions. Is variation within genes responsible for the apparent variation in G+C content among genes? If not, what are the relative strengths of the among-gene and within-gene sources of variation? Kliman and Hey (1994) found a significant positive correlation between intron G+C content and third codon position G+C content across genes, indicating significant variation in intron (and third position) G+C content among genes. However, this correlation could be caused by variation within a gene. One-way analysis of variance (ANOVA) on the intron G+C content data finds significant variance among genes that could not be explained by variance within genes ( $F_{85,282} = 1.906$ ,  $P < 10^{-4}$ ). This is consistent with the finding of substantial variation in composition among large fragments of the *D. melanogaster* genome (Carulli et al. 1993).

The relative strengths of the among-gene and within-gene factors can be assessed by estimating their contributions to the variance in intron G+C content. A nested ANOVA showed, surprisingly, that the within-gene variance in percentage G+C (33.17%) is larger than the among-gene variance (16.75%).

### G+C Content Correlations Within Genes

Since there is substantial variation in intron G+C content within *D. melanogaster* genes, it is of interest to see if

**Table 2.** Within-gene correlations<sup>a</sup>

	<i>N</i>	<i>r</i> <sub>within</sub>	<i>P</i>
A			
Codon position			
First	86	-0.085	0.960
Second	86	-0.076	0.918
Third	86	0.177	0.0005
B			
Number of introns			
1	47	0.214	0.032
2	21	0.194	0.064
3	10	0.224	0.054
4	1	0.309	0.227
5	2	0.436	0.036
6	1	-0.388	0.883
7	1	0.051	0.431
11	1	0.224	0.158
13	1	0.371	0.032
14	1	-0.011	0.524
≥2	39	0.176	0.002

<sup>a</sup> *P* is the proportion of 100,000 randomized data sets which gave an  $r_{\text{within}}$  value greater than that calculated from the original data. (A) Correlations between intron G+C content and either first, second, or third codon position G+C content. The correlation involving the third codon position is significant after applying the sequential Bonferroni correction for multiple comparisons (Rice 1989). (B) Correlations between intron and codon third position G+C content for subsets of genes classified by intron number.

similar variation is apparent within exons. Table 2A shows that this is indeed the case; the intron G+C content is highly correlated to the third codon position G+C content but not to that of the first or second codon positions. However, because the correlation coefficient tends to be dominated by genes with many introns, it is possible that compositional heterogeneity is limited to a few highly segmented genes. To address this, we repeated the analysis, grouping genes by number of introns. Although intron G+C content is only significantly correlated [before correction for multiple comparisons (Rice 1989)] to third position G+C content for the genes with one intron and for 2 other analyses, 8 of the 10 analyses gave positive correlations, with a few approaching statistical significance. Statistical significance does not appear to depend on the inclusion of genes with one intron, as the within-gene correlation remains significantly positive when these genes are removed. Further, if we calculate the correlation coefficient for each gene individually, significantly more than half of the genes show a positive correlation between intron and third codon position G+C content (53 of 86;  $z = 2.16$ ,  $P < 0.05$ ). The results indicate that the correlation between intron and third position G+C content within genes is widespread.

### Variation Along Genes

Variation in composition within genes might take a relatively simple form, such as an increase or decrease in

**Table 3.** G+C content and codon usage along genes with introns<sup>a</sup>

Variable	$b_{\text{within}}$	$P$
A		
Intron G+C	$-1.49 \times 10^{-5}$	0.000
First codon position G+C	$1.67 \times 10^{-6}$	0.338
Second codon position G+C	$1.29 \times 10^{-6}$	0.428
Third codon position G+C	$-7.54 \times 10^{-6}$	0.000
Preferred codon usage	$-9.43 \times 10^{-6}$	0.000
B		
Intron G+C	-0.0110	0.000
First codon position G+C	0.0029	0.724
Second codon position G+C	0.0199	0.016
Third codon position G+C	-0.0297	0.000
Preferred codon usage	-0.0321	0.000

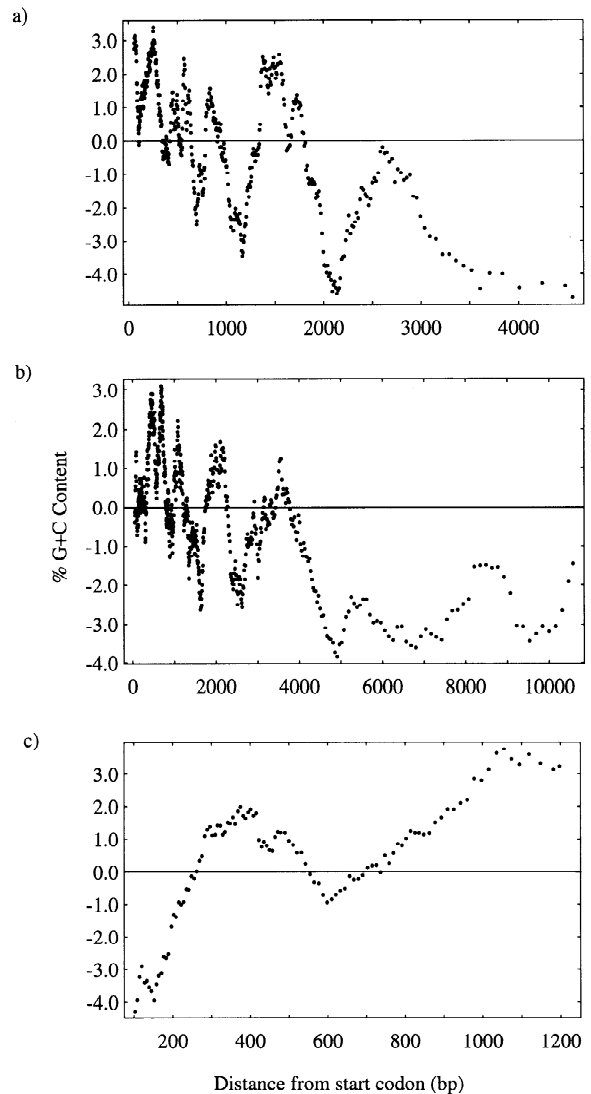
<sup>a</sup>  $P$  is the proportion of 1,000 randomized data sets that gave an absolute value of  $b_{\text{within}}$  greater than the absolute value of  $b_{\text{within}}$  calculated from the original data. Analyses did not include the first codon, as this is invariant in all eukaryotic genes. (A) Distance was measured as the number of bases past the start codon. (B) Distance was measured as the relative position between the start and the stop codons.

G+C content along the length of the gene. Furthermore, if introns and third codon positions shared a similar trend, this may explain the within-gene correlations between the two. To address this, we performed least-squares regression analyses of intron and exon G+C content along genes, after factoring out the difference among genes.

Slopes of lines through the data and Monte Carlo estimates of statistical significance are given in Table 3A. In genes with introns, there is a significant decline in intron and third codon position G+C content but no trend in either first or second position G+C content. The slopes equate to declines in G+C content of  $\sim 1.5$  and  $0.75\%$  per 1,000 base pairs (bp) for introns and third positions, respectively. This decline is evident when the data are plotted in such a way as to remove the differences between genes (Figs. 1a and b).

Because long genes are disproportionately weighted in regression analyses using absolute distance (unscaled analysis) from the start codon, we repeated the analyses using relative distance (scaled analysis) between the start and the stop codon. The results are qualitatively similar to those from the unscaled analysis (Table 3B); there is a significant decline in both intron and third codon position G+C content. The decline in intron G+C content appears to be particularly widespread, since significantly more than half the genes (54 of 86;  $z = 2.37$ ,  $P < 0.02$ ) show a decline when analyzed individually. The generality of the decline in third position G+C is less evident, with negative slopes in just over half (44 of 86) of the genes.

A few genes (e.g., *Arrl*, *His2AvD*, *LanA*, *rdgC*, and *ry*) contributed strongly to the intron G+C content regression, even after scaling the distances along the gene. However, it should be noted that only *LanA* and *ry* were large contributors to the negative regressions for third



**Fig. 1.** Trends in G+C content along genes. Figures show the average G+C content corrected for differences among genes plotted against the distance from the first base of the start codon for (a) intron and (b) third position G+C content for genes with introns and (c) third position G+C content for genes without introns. The gene mean G+C content (intron or exon, as appropriate) was subtracted from each point in each gene. The data were then ordered, across genes, according to the distance of each point from the start codon, and moving averages of 2,000 points calculated. The window for the moving average was sequentially advanced by at least 50 points; since there are many ties (e.g., for most genes, the third position of the second codon is 5 bp from the first base of the coding sequence), the window was advanced by a variable number of base pairs.

position G+C and preferred codon usage. Thus, the within-gene correlations between intron and third position G+C content do not appear to arise from a small number of genes sharing similar along-gene trends in intron and third position G+C content.

Genes with no introns show a strikingly different trend in base composition. There is a highly significant increase in third position G+C content in these genes. Increases in first and second position G+C content are not significant (Table 3). The increase in third position

**Table 4.** G+C content and codon usage along genes lacking introns

Variable	$b_{\text{within}}$	$P$
A		
First codon position G+C	$8.05 \times 10^{-6}$	0.573
Second codon position G+C	$2.11 \times 10^{-6}$	0.886
Third codon position G+C	$6.06 \times 10^{-5}$	0.000
Preferred codon usage	$7.15 \times 10^{-5}$	0.000
B		
First codon position G+C	0.0146	0.453
Second codon position G+C	0.0077	0.683
Third codon position G+C	0.0612	0.000
Preferred codon usage	0.0726	0.000

<sup>a</sup> See Table 3, footnote a, for details.

G+C content, depicted in Fig. 1c, is equivalent to ~6% per 1,000 bp.

In *Drosophila*, the genes with no introns tend to be shorter overall. It is, therefore, possible that a general increase in G+C content at the start of genes is obscured in the genes with introns by the decrease farther along the gene. When the analysis is limited to the first 350 or 500 codons of genes with introns, there is no evidence of an increasing G+C trend. However, when we limit the analysis to the first 200 codons, a significant increase in third codon position G+C slope is detected (for scaled distance  $b_{\text{within}} = 0.055 \pm 0.026$ ,  $P = 0.024$ , for genes with introns and  $b_{\text{within}} = 0.108 \pm 0.041$ ,  $P = 0.008$ , for genes without). The two slopes are not significantly different (Sokal and Rohlf 1995, p. 493), indicating similar base composition trends at the start of coding regions, regardless of whether or not they have introns. This increase is detectable in the majority of genes; among genes with introns, 45 of 65 had positive third codon position G+C slopes over the first 200 codons, compared to 14 of 18 genes lacking introns.

Selection is believed to have acted upon synonymous codon use in *D. melanogaster* (Shields et al. 1988; Kliman and Hey 1993, 1994; Akashi 1994, 1995). Since all of the preferred codons in *D. melanogaster* are C- or G-ending [16 C-ending and 6 G-ending (Akashi 1995)], trends in third codon position G+C content are not independent of trends in preferred codon usage. As expected, there is a significant decrease in preferred codon usage toward the 3' end of genes with introns and a significant increase in genes without introns (Tables 3 and 4). Furthermore, the significant increase in G+C content in the first 200 codons of all genes is reflected in a significant increase in preferred codon usage for both classes of genes (for scaled distance  $b_{\text{within}} = 0.08 \pm 0.026$ ,  $P = 0.001$ , for genes with introns and  $b_{\text{within}} = 0.152 \pm 0.045$ ,  $P = 0.000$ , for genes without introns). Again, the two slopes are not significantly different.

## Discussion

We have shown that there is highly significant heterogeneity in intron base composition within *D. melanogaster*

genes, confirming the results of previous work (Kliman and Hey 1994). This appears to be due, in part, to a decline in intron G+C content along many genes. There is also an overall decline in third codon position G+C content along genes with introns. Not surprisingly, there is a positive correlation between intron and third position G+C content within genes. However, at the start of genes there appears to be an increase in third position G+C content over the first few hundred codons; this is particularly evident in genes without introns.

A central question regarding base composition heterogeneity is whether the composition variation within genes is a "snapshot" of the variation at a much larger scale. The trends along individual genes, of the order of 2 kb, may reflect larger-scale trends. Several lines of evidence challenge this. First, ANOVA indicated highly significant variation among genes not explained by variation within genes. Still, some caution should be exercised, since the ANOVA does not consider the structure of the intron variation (i.e., the apparent decline in G+C content along genes). Second, the general decline in G+C content along intron-bearing genes would be difficult to achieve unless the genes were nonrandomly positioned with respect to the large-scale variation in G+C content. Third, the decline in G+C content within a gene is too great to be consistent with the variation in G+C content seen among chromosomal segments. Carulli et al. (1993) found that large blocks of DNA varied from 36.9 to 50.9% G+C, with a variance of 8.07. We find at least as much variance within a gene as they found among blocks.

Base composition heterogeneity can arise from variation in mutational patterns, selection, and biased gene conversion. Variation in mutational patterns is highly consistent with our observations; similar variation in G+C content is observed in introns and exons, and the variation appears to be structured in a way one might expect of mutation (i.e., a simple increase or decrease in G+C content). Large-scale variation in composition has been attributed to differences in replication time and changing conditions within the cell (Wolfe et al. 1989). This is an unlikely explanation in the present case, because either the conditions under which replication occurred would have to change very rapidly or many genes would have to span boundaries between replicons. While there is evidence that adjacent replicons in *Drosophila* do not form in synchrony, unlike in vertebrates, replicons are thought to be much larger than individual genes (Steinemann 1981). Therefore, variation within genes in base composition is not likely to be generated by replication time effects. Nor is it likely to be generated by variation in the pattern of methylation since methylation appears to be absent from *Drosophila* DNA (Bird 1987). An alternative explanation is variation in the efficiency of DNA repair (Filipki 1987, 1988; Eyre-Walker 1994). Various types of DNA repair are known to vary in their

efficiency over the scale of a gene (Bohr et al. 1987; Boulikas 1992) and this could generate variation in the pattern of mutation (Eyre-Walker 1994).

Selection could also be responsible for the variation in base composition within genes. For example, selection seems to constrain intron secondary structure (Stephan and Kirby 1993; Schaeffer and Miller 1993; Leicht et al. 1995; Kirby et al. 1995). Since the propensity to form, and the stability of, secondary structures, depends upon the base composition, selection could generate trends in base composition along genes.

It has been suggested that biased gene conversion might explain the relationship between recombination rate and G+C content in humans (Holmquist 1992; Eyre-Walker 1993). This is an unlikely explanation for the variation in composition along genes. Kliman and Hey (1993) found no evidence that intron G+C content among genes was related to recombination frequency in *D. melanogaster*. Furthermore, gene conversion could be responsible for our observations only if the variance in recombination frequency along a gene exceeded the variance among genes. We are unaware of any evidence for variation in recombination frequency along genes.

Although both intron and third position G+C contents are correlated with each other and show the same overall pattern along genes, they do not necessarily have the same cause. It is intriguing that the declines in intron and third position G+C content seem to be due mainly to G and C, respectively. In introns, the G scaled distance regression slope was  $-0.0858$  ( $P = 0.000$ ), while the C slope was  $-0.0243$  ( $P = 0.028$ ). For third codon positions, the G slope was  $-0.0099$  ( $P = 0.204$ ) and the C slope was  $-0.0197$  ( $P = 0.015$ ). This difference may reflect the fact that selection is thought to act, or to have acted, upon synonymous codon use in *D. melanogaster* (Shields et al. 1988; Kliman and Hey 1993, 1994; Akashi 1994, 1995). Since 16 of 22 preferred codons are C-ending (Akashi 1995), the decline in C along genes may reflect a decline in the strength of selection on synonymous codon bias or conflicting selection pressures.

Although there is an overall decline in third position G+C content along genes, there appears to be an increase in the first few hundred codons of most genes. This is particularly evident in the genes which have no introns. In fact, there is no evidence of a decline farther along the genes with no introns; this may reflect their length or some other qualitative difference. The increase in G+C content at the start of genes could itself be caused by mutation, selection, or biased gene conversion. Unfortunately it is not possible to examine trends in introns near the start of genes since there are insufficient data. However, the increase in third position G+C content and preferred codon use of the first 200 codons is positively correlated with the overall synonymous codon bias in the gene as measured by the codon adaptation index (Sharp and Li 1987; Sharp et al. 1992; Kliman and Hey 1994)

(third position G+C slope vs CAI,  $r = 0.333$ , 81 df,  $P = 0.002$ ; preferred codon slope vs CAI,  $r = 0.309$ , 80 df,  $P = 0.005$ ). This is reminiscent of the pattern seen in *Escherichia coli*, where the increase in codon bias at the start of genes is most evident for highly biased genes (Eyre-Walker and Bulmer 1993). It is most consistent with a model in which there are conflicting selection pressures acting upon synonymous codon use at the start of the gene; i.e., there is selection to use the codons preferred in the middle of the gene, but this selection is overcome at some sites by other conflicting selection pressure. For example, there might be selection to avoid secondary structure at the start of the mRNA. If the increase in G+C content was associated with mutation or biased gene conversion, one would expect the increase to be most evident in the low biased genes.

If conflicting selection pressures are responsible for lower codon bias at the start of *D. melanogaster* genes, we might expect the rate of synonymous substitution to be lower at the start of the gene than in the middle, as we see in *E. coli* (Eyre-Walker and Bulmer 1993). Analysis confirms that this is the case; the level of synonymous divergence between *D. melanogaster* and *D. pseudoobscura* (corrected for differences in average synonymous substitution rate between genes) increases significantly over the first 200 codons, by approximately 0.10 synonymous substitutions per site (Spearman's rank correlation coefficient  $r_s = 0.255$ ,  $P = 0.0003$ ). Over the next 200 codons the correlation is positive but nonsignificant ( $r = 0.070$ ,  $P = 0.323$ ), and then over the remaining codons it is negative ( $r_s = -0.029$ ,  $P = 0.374$ ). It should be noted that there are few data available past 200 codons so the last two results should be treated with caution.

Although the reasons for the compositional trends within genes remain obscure, our preliminary analysis of *C. elegans* genes suggests that other eukaryotes may show similar trends. In addition to an increase in preferred codon usage at the start of genes, we find a significant decline in intron G+C content (unpublished results). With the accumulation of DNA sequence data for *S. cerevisiae*, as well as substantial data for *Dictyostelium discoideum*, it will be of great interest to see if trends in codon usage and base composition are general phenomena in eukaryotes.

*Acknowledgments.* We thank Hiroshi Akashi and Brandon Gaut for discussion and helpful comments on the manuscript and Etsuko Moriyama for providing the aligned sequences used in this study. This study was supported by the Radford University College of Arts and Sciences and by a grant to R.M.K. from the Jeffress Memorial Trust. A.E.W. is a Royal Society University Research Fellow.

## 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
- Bernardi G (1989) The isochore organization of the human genome. *Annu Rev Genet* 23:637–661
- Bernardi G, Olofsson B, Filipiski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228:953–958
- Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. *Trends Genet* 3:342–347
- Bohr VA, Philips DH, Hanawalt PC (1987) Heterogeneous DNA damage and repair in the mammalian genome. *Cancer Res* 47:6426–6436
- Boulikas T (1992) The evolutionary consequences of nonrandom damage and repair of chromatin domains. *J Mol Evol* 35:156–180
- Carulli JP, Krane DE, Hartl DL, Ochman H (1993) Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* genome. *Genetics* 134:837–845
- Deschavanne P, Filipiski J (1995) Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. *Nucleic Acids Res* 23:1350–1353
- Dujon B, et al. (1994) Complete DNA sequence of yeast chromosome XI. *Nature* 369:371–378
- Eyre-Walker A (1993) Recombination and mammalian genome evolution. *Proc R Soc Lond B* 252:237–243
- Eyre-Walker A (1994) DNA mismatch repair and synonymous codon evolution in mammals. *Mol Biol Evol* 11:88–98
- Eyre-Walker A (1996) The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. *J Mol Evol* 42:73–78
- Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* 21:4599–4603
- Feldmann H, et al. (1994) Complete DNA sequence of yeast chromosome II. *EMBO J* 13:5795–5809
- Filipiski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett* 217:184–186
- Filipiski J (1988) Why the rate of silent substitution is variable within a vertebrate's genome. *J Theor Biol* 134:159–164
- Gardiner K, Aissani B, Bernardi G (1990) A compositional map of human chromosome 21. *EMBO J* 9:1853–1858
- Holmquist GP (1992) Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet* 40:151–173
- Kirby DA, Muse SV, Stephan W (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci USA* 92:9047–9051
- 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 (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* 137:1049–1056
- Leicht GG, Muse SV, Hanczyc M, Clark AG (1995) Constraints on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila*. *Genetics* 139:299–308
- Lewis J, Bird A (1991) DNA methylation and chromatin structure. *FEBS Lett* 285:155–159
- Oliver SG, et al. (1992) Complete DNA sequence of yeast chromosome III. *Nature* 357:38–46
- Pilia G, Little RD, Aissani B, Bernardi G, Schlessinger D (1993) Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics* 17:456–462
- Rice WR (1989) Analyzing tables of statistical tests. *Evolution* 43:223–225
- Schaeffer SW, Miller EL (1993) Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* 135:541–552
- Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon bias, and its potential application. *Nucleic Acids Res* 15:1281–1295
- Sharp PM, Lloyd AT (1993) Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic Acids Res* 21:179–183
- Sharp PM, Burgess CJ, Lloyd AT, Mitchell KJ (1992) Selective use of termination codons and variations in codon choice. In: Hatfield DL, Lee GJ, Pirtle RM (eds) *Transfer RNA in protein synthesis*. CRC Press, Boca Raton, FL, pp 397–425
- Shields D, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* 5:704–716
- Sokal RR, Rohlf FJ (1995) *Biometry*. Freeman, New York
- Steinemann M (1981) Chromosomal replication in *Drosophila virilis*. III. Organization of active origins in the highly polytene salivary gland cells. *Chromosoma* 82:289–307
- Stephan W, Kirby DA (1993) RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. *Genetics* 135:97–103
- Wolfe K, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285