Compositional Heterogeneity and Patterns of Molecular Evolution in the Drosophila Genome

John P. Carulli,* Dan E. Krane,† Daniel L. Hartl†,[1] and Howard Ochman‡

*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, †Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, and ‡Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received October 28, 1992
Accepted for publication March 25, 1993

ABSTRACT

The rates and patterns of molecular evolution in many eukaryotic organisms have been shown to be influenced by the compartmentalization of their genomes into fractions of distinct base composition and mutational properties. We have examined the Drosophila genome to explore relationships between the nucleotide content of large chromosomal segments and the base composition and rate of evolution of genes within those segments. Direct determination of the G + C contents of yeast artificial chromosome clones containing inserts of *Drosophila melanogaster* DNA ranging from 140–340 kb revealed significant heterogeneity in base composition. The G + C content of the large segments studied ranged from 36.9% G + C for a clone containing the *hunchback* locus in polytene region 85, to 50.9% G + C for a clone that includes the *rosy* region in polytene region 87. Unlike other organisms, however, there was no significant correlation between the base composition of large chromosomal regions and the base composition at fourfold degenerate nucleotide sites of genes encompassed within those regions. Despite the situation seen in mammals, there was also no significant association between base composition and rate of nucleotide substitution. These results suggest that nucleotide sequence evolution in Drosophila differs from that of many vertebrates and does not reflect distinct mutational biases, as a function of base composition, in different genomic regions. Significant negative correlations between codon-usage bias and rates of synonymous site divergence, however, provide strong support for an argument that selection among alternative codons may be a major contributor to variability in evolutionary rates within Drosophila genomes.

MUCH of the information concerning the organization and evolution of the Drosophila genome has come from two sources: the banding patterns of polytene chromosomes (ASHBURNER 1989) and the examination of genetic polymorphisms, particularly those revealed by nucleotide sequencing and protein electrophoresis (LEWONTIN 1985, 1991). Both types of analysis have revealed a mosaic structure of Drosophila chromosomes. The euchromatic portion of the Drosophila genome comprises approximately 5000 polytene chromosome bands (SORSA 1988), and the presence of these alternately dark- and light-staining regions indicates that segments of the chromosome are sufficiently heterogeneous to produce the observed patterns of banding. In addition to cytological features, rates of nucleotide sequence evolution display variation over the chromosome, and adjacent regions may exhibit different patterns of change (MARTIN and MEYEROWITZ 1988). Furthermore, comparisons of homologous sequences from several species of Drosophila show that substitution rates at synonymous sites—those changes that do not alter the amino acid composition of the protein and are typically thought to reflect neutral mutations—vary considerably among nuclear genes (RILEY 1989; SHARP and LI 1989).

In mammals, the observed heterogeneity in chromosome structure and evolutionary rates among genes has been attributed to large-scale differences in nucleotide content over the genome (BERNARDI 1989; WOLFE, SHARP and LI 1989). Several lines of evidence suggest that mammalian genomes are compartmentalized into extensive regions of relatively homogeneous base composition, called "isochores" (BERNARDI *et al.* 1985). Structure at the level of base composition within the genomes of birds and mammals was originally detected by the heterogeneous distribution of DNA fragments in buoyant density gradients (SCHILDKRAUT and MAIO 1969; BERNARDI *et al.* 1985). The metaphase banding patterns of mammals may reflect this physical organization of chromosomes (CUNY *et al.* 1981; IKEMURA and AOTA 1988; BERNARDI 1989) in that giemsa-stained (G) bands of chromosomes correspond to AT-rich, late-replicating DNA whereas the reverse (R) bands are GC-rich, early replicating and contain a majority of active genes (GOLDMAN *et al.* 1984; BICKMORE and SUMNER 1989). Statistical anal-

[1] To whom correspondence should be addressed.

yses of nucleotide sequences also revealed an underlying base compositional structuring to mammalian genomes. For example, the synonymous sites, introns and DNA flanking a given gene all have similar base compositions, indicating that isochores may extend over regions many hundreds of kilobases long. Furthermore, the rates of nucleotide substitutions at silent sites vary with the G + C content of genes and noncoding DNA (FILIPSKI 1987; WOLFE, SHARP, and LI 1989; IKEMURA, WADA and AOTA 1990), and regions of different base content may also be subject to differences in rates of recombination (HARDISON *et al.* 1991) as well as different frequencies of stable insertion of repetitive elements (BOYLE, BALLARD and WARD 1990, Hardison *et al.* 1991).

Despite the extensive use of Drosophila in genetic studies, the striking banding patterns of Drosophila polytene chromosomes and the wide variation in silent substitution rates among genes, relatively little is known about the broad-scale structure of the Drosophila genome with respect to base composition. Nuclear DNA (excluding satellite sequences) from *Drosophila melanogaster* produces homogeneous banding profiles in buoyant density gradients, indicating an absence of sharply defined regions whose base compositions widely differ (THIERY, MACAYA and BERNARDI 1976). In addition, there is no significant association between the G + C contents of introns and silent sites in *D. melanogaster* (SHIELDS *et al.* 1988; MORIYAMA and HARTL 1993). It is now possible, however, to perform more detailed analyses of the physical structure of the Drosophila genome due to the advent of new techniques including yeast artificial chromosome cloning (BURKE *et al.* 1987; GARZA *et al.* 1989), pulsed field gel electrophoresis (CHU, VOLLRATH and DAVIS 1986) and base ratio analysis with thin-layer chromatography (KRANE, HARTL and OCHMAN 1991).

In an attempt to determine the extent to which DNA sequence variation is correlated with the organization of chromosomes, the present study examines the nucleotide content across large regions of the Drosophila genome. Much of the *D. melanogaster* genome has been cloned in the form of yeast artificial chromosomes (GARZA *et al.* 1989) whose locations have been mapped by *in situ* hybridizations (AJIOKA *et al.* 1991; HARTL 1992). The availability of these clones makes it possible to determine whether the Drosophila genome is structured on a scale not resolved by conventional analyses. The base composition of yeast artificial chromosome clones containing large regions of *D. melanogaster* DNA from 18 unique euchromatic locations was therefore determined. In addition, by focusing on chromosomal segments that encompass genes whose nucleotide sequences are available from two or more species of Drosophila, we have been able to assess the relationship between regional base composition and rates of molecular evolution across larger regions than had previously been possible.

MATERIALS AND METHODS

Isolation of yeast artificial chromosome (YAC) clones: YAC clones were selected with respect to their chromosomal location, insert size and genetic content. Many of the YACs included in this study contain genes whose nucleotide sequences have been determined for *D. melanogaster* and at least one other drosophilid, usually *Drosophila pseudoobscura*. Several YACs were isolated by screening 3456 yeast colonies stamped in ordered arrays on nylon filters. Probes for screening the library included cloned *D. melanogaster* genes and sequences amplified from *D. melanogaster* genomic DNA by the polymerase chain reaction (SAIKI *et al.* 1988). Oligonucleotide primers were based on published sequences (Genbank release 69) and designed to amplify regions of several hundred base pairs from the corresponding genes. Probes were radiolabeled by the method of FEINBERG and VOGELSTEIN (1983), and hybridization was carried out at 65° for 16 hr in a solution of 1 M NaCl, 1% sodium lauryl sulfate (SDS) and 10% dextran sulfate. Prior to autoradiography, filters were washed in two changes of 2 × SSC (20 × SSC is 3 M NaCl and 0.3 M sodium citrate, pH 7.0) for 5 min each at room temperature, followed by two changes of 2 × SSC containing 1% SDS for 30 min each at 65°, and a final rinse in 0.1 × SSC for 30 min at room temperature.

Remaining YACs included in the study were identified by screening smaller numbers of clones that were previously mapped cytologically (AJIOKA *et al.* 1991; HARTL 1992). Either the polymerase chain reaction or Southern hybridization (SOUTHERN 1975) was employed to assay the presence of *D. melanogaster* genes within the YAC inserts. PCR-based assays were performed directly on colonies of yeast harboring artificial chromosomes to confirm the genetic content of the Drosophila inserts. Alternatively, the chromosomal DNAs of yeast clones containing YACs from a given cytological region were electrophoresed through agarose gels and transferred to nylon membranes. Probes were prepared and radiolabeled as described above. Hybridization proceeded for 16 hr at 65° in phosphate buffer (0.5 M NaCl, 0.1 M NaH_2PO_4, 5 mM EDTA) supplemented with 0.2% Sarkosyl, and filters were washed for a total of 30 min in five changes of 10 mM Tris, 1 mM EDTA (pH 8.0) prior to autoradiography. The map position of each YAC was confirmed by *in situ* hybridization of biotin-labeled DNA to preparations of salivary gland polytene chromosomes from *D. melanogaster* (AJIOKA *et al.* 1991).

Base composition of YAC clones: YACs containing fragments of DNA from *D. melanogaster* Oregon R were propagated in yeast strain AB1380 (GARZA *et al.* 1989). Intact chromosomal DNA from cells grown in 5-ml cultures was prepared in agarose plugs, and YACs were separated from the complement of yeast chromosomes by pulsed-field gel electrophoresis (CARLE and OLSON 1984; CHU, VOLLRATH and DAVIS 1986). DNA corresponding to the YAC was excised from the gel and extracted from agarose by the glass-powder procedure (GeneClean, Bio 101). The base composition of each YAC was determined by base ratio analysis with TLC (KRANE, HARTL and OCHMAN *et al.* 1991), and the final values reflect the average for at least three replicates of each sample.

Data analysis: Estimates of the numbers of synonymous and nonsynonymous base substitutions in comparisons between homologous sequences were calculated by the method of LI, WU and LUO (1985), and the base compositions of published nucleotide sequences were determined by simple tabulation. Values of F_{op}, a measure of the relative occurrence of optimal codons in a protein-coding region, were computed by the method of IKEMURA and OZECKI (1983), using the codon usage table for *D. melanogaster* sequences compiled and provided by PAUL SHARP (personal communication). A second measure of codon-usage bias, χ^2/L (SHIELDS *et al.* 1988), was calculated using software provided by ETSUKO MORIYAMA (MORIYAMA and HARTL 1993). For the analysis of the relation between codon-usage bias and rate of sequence divergence, the average of the *D. melanogaster* and *D. pseudoobscura* F_{op} or χ^2/L values for each gene were used. Many of the statistical analyses were performed with the StatView statistical package. For the analysis of variance among the G + C contents for the YAC clones, the square roots of the G + C percentages were arcsine transformed to approximate a normal distribution (SOKAL and ROHLF 1969).

RESULTS

Nucleotide composition varies within the Drosophila genome: The G + C contents of 18 YAC clones, which range in size from 140–340 kb, vary from 36.9–50.9% (Table 1, Figure 1). Analysis of variance in base composition among three to six replicates for each of the 18 YAC clones demonstrates that there is significantly more variation among the YAC clones than there is among replicates for a given YAC ($F_{17,48} = 44.96$, $P < 0.0001$; 91.5% of variance among clones, 8.5% of variance between replicates of clones), even when the highest and lowest G + C percentages are removed from the analysis ($F_{15,43} = 24.36$, $P < 0.0001$; 83.0% of variance among clones, 17.0% of variance between replicates of clones). Despite the diversity in base composition among the cloned inserts of Drosophila DNA, the mean G + C content of all the YACs was 43.9%, which compares favorably with the base composition of 43.0% (ASHBURNER 1989) determined for the entire *D. melanogaster* genome. In some regions of the genome, nucleotide composition varies considerably, even over a span of a few megabases. In fact, the two extreme values—for YAC clones N02-20 (36.9% G + C) and DY280 (50.9% G + C)—occur within two major polytene divisions on the right arm of the third chromosome. However, in other regions of the genome, such as polytene divisions 89, 91 and 92 on chromosome arm 3R, broad-scale base content is consistent over distances of several megabases. Superimposed on this substructure, there is a tendency toward increasing G + C content with distance from the chromocenter ($r = 0.373$, $P = 0.134$; with the extreme values removed $r = 0.463$, $P = 0.08$).

The G + C contents of the YACs studied were also compared with (1) the G + C contents of coding sequences of the genes they contain, and (2) the G + C content at fourfold degenerate nucleotide sites in those genes. Although there is considerable heterogeneity in the nucleotide composition of YACs and sequenced genes, there is no significant correlation between the nucleotide composition on the broad scale of hundreds of kilobases and the nucleotide composition of genes embedded within those large regions ($r = -0.15$, $P = 0.5$; Figure 2). There is also no apparent association between the G + C contents of the YACs and the G + C contents at fourfold degenerate nucleotide sites in the genes ($r = -0.162$, $P = 0.46$, Figure 2). For YACs that included multiple genes, the G + C contents at fourfold degenerate nucleotide sites ranged from 62.8–79.8% in N01-65 and from 60.0–74.1% in N03-42 (Table 1).

Nucleotide composition and synonymous site evolution: Analysis of the rates of evolution for loci contained within the YACs studied was restricted to loci whose nucleotide sequences have been determined for both *D. melanogaster* and a species from the *obscura* group. These comparisons require no assumptions about divergence times, which would have been necessary if comparisons among multiple species with differing degrees of relatedness were included. Synonymous site divergence rates (K_s) for the genes used in this study ranged from 0.40 for *Rh1* to 1.45 for *pcp* (Table 2). A tendency for the rate of sequence divergence to increase with the G + C content was found (Figure 3), but the association between the two factors is not statistically significant ($r = 0.375$, $P = 0.24$). When the K_s for the most rapidly evolving gene, *pcp*, is removed from the analysis, the correlation becomes stronger but is still not significant ($r = 0.501$, $P = 0.19$).

Nucleotide composition and codon-usage bias: Factors in addition to base composition, such as codon-usage bias, may influence the rate of synonymous site evolution in Drosophila. For the genes examined here, codon-usage bias is independent of the G + C content of the YACs containing them ($r = 0.044$, $P = 0.85$). As previously determined for a less comprehensive set of Drosophila sequences (SHARP and LI 1989; SHIELDS *et al.* 1988), there is a significant negative correlation between codon-usage bias and the rate of nucleotide substitution (F_{op} *vs.* K_s: $r = 0.595$, $P = 0.02$; χ^2/L *vs.* K_s: $r = 0.605$, $P = 0.008$; Figure 4). In Drosophila, as in bacteria (SHARP and LI 1987), genes with strong codon-usage bias fix mutations at a slower rate, indicating that codon-usage bias exerts significant constraints on the rate of synonymous site evolution in these organisms.

DISCUSSION

Base composition varies with chromosomal position in *D. melanogaster*: YAC clones containing 140–340 kb of DNA from unique, euchromatic portions of the *D. melanogaster* genome exhibit significant het-

TABLE 1

Size, genomic location, base composition and genetic content of *D. melanogaster* YAC clones

YAC No.	Size (kb)	Gene[a]	Map Position[b]	Distance[c] (mb)	%G + C YAC (SE)[d]	%G + C$_4$[c]
DY396	160	*sevenless*	10A1-2 (X)	10.9	42.7 (0.3)	66.1
N01-65	210	*Gart*	27C (2L)	12.3	45.6 (0.7)	62.8
N01-65	210	*pcp*	27C (2L)	12.3	45.6 (0.7)	79.8
N05-59	200	*spalt*	33A2 (2L)	7.3	45.5 (1.0)	26.0
N02-27	250	*Adh*	35B3 (2L)	5.1	41.2 (0.2)	80.4
DY132	200	*Ddc*	37C1-2 (2L)	2.7	45.5 (0.2)	71.6
N05-96	210	*Lcp*	44D (2R)	3.7	44.1 (0.3)	65.3
N03-81	150	*engrailed*	48A3-4 (2R)	7.1	40.4 (0.6)	76.1
DY177	200	*hsp82*	63B-C (3L)	18.4	44.0 (0.5)	74.2
DY178	190	*chorion s15*	66D11-15 (3L)	13.7	45.2 (0.7)	56.5
DY178	190	*chorion s19*	66D11-15 (3L)	13.7	45.2 (0.7)	59.3
N03-42	340	*Sod*	68A8-9 (3L)	11.7	42.1 (0.7)	74.1
N03-42	340	*sgs-8*	68C3-5 (3L)	11.5	42.1 (0.7)	60.0
N03-42	340	*sgs-7*	68C3-5 (3L)	11.5	42.1 (0.7)	60.6
N03-42	340	*sgs-3*	68C3-5 (3L)	11.5	42.1 (0.7)	66.5
N02-20	140	*hunchback*	85A3-B1 (3R)	3.7	36.9 (0.6)	70.5
DY280	170	*rosy*	87D8-12 (3R)	7.6	50.9 (0.4)	65.0
DY338	190	*Ubx*	89E (3R)	10.9	43.0 (0.2)	75.5
N03-88	170	*Rh2*	91D1-2 (3R)	12.4	43.4 (0.2)	69.4
N10-42	180	*Rh1*	92B6-7 (3R)	13.3	43.3 (0.5)	82.0
N11-84	140	*Rh3*	92D1 (3R)	13.5	43.2 (0.2)	80.9
N04-56	260	*rough*	97D5-7 (3R)	20.2	44.2 (0.4)	69.3
N04-74	260	*rp49*	99D (3R)	22.5	49.3 (0.3)	78.8

[a] Abbreviations, accession numbers and citations for sequences are as follows: *sevenless* (J03158; BASLER and HAFEN 1988); *Gart* (J02527; HENIKOFF *et al.* 1986); *pcp* = pupal cuticle protein (J02527; HENIKOFF *et al.* 1986); *spalt* (X57474; FREI *et al.* 1988); *Adh* = alcohol dehydrogenase (M19264; BODMER and ASHBURNER, 1984); *Ddc* = dopa decarboxylase (X04661; EVELETH *et al.* 1986); *Lcp* = larval cuticle protein (J01080; SYNDER *et al.* 1982); *engrailed* (M10017; POOLE *et al.* 1985); *hsp82* = heat shock protein (X03810; BLACKMAN and MESELSON 1986); chorion proteins *s15* and *s19* (X02497; WONG *et al.* 1985); *Sod* = superoxide dismutase (X13780; KWIATOWSKI, PATEL and AYALA 1989); *sgs-3*, *sgs-7*, *sgs-8* = salivary gland secretion proteins (X01918; GARFINKEL PRUITT and MEYROWITZ 1983); *hunchback* (Y00274) (TAUTZ *et al.* 1987); *rosy* = xanthine dehydrogenase (Y00308; LEE *et al.* 1987; KEITH *et al.* 1987); *Ubx* = *ultrabithorax* (K01963; WILDE and AKAM 1987); *Rh2* = ocellar opsin protein (M12896; COWMAN, ZUKER and RUBIN 1986); *Rh1* = *ninaE* = R 1-6 opsin protein (K02315; O'TOUSA *et al.* 1985; ZUKER, COWMAN and RUBIN 1985); *Rh3* = R7 opsin protein (M17730; ZUKER *et al.* 1987); *rough* (M23629; TOMLINSON, KIMMEL and RUBIN 1988); *rp49* = ribosomal protein 49 (X00848; KONSUWAN *et al.* 1985).

[b] Polytene band designations are followed by chromosome arm location in parentheses.

[c] Numbers reflect the distance, in megabases, from the chromocenter.

[d] SE is the standard error of the mean G + C percent determined from three to six replicates for each YAC.

[e] Base composition at fourfold degenerate nucleotide sites.

erogeneity in regional base composition, indicating that the Drosophila genome is more heterogeneous than studies of banding profiles in buoyant density gradients have suggested. In regard to this level of heterogeneity, the structure of the Drosophila genome may be analogous to the structure of mammalian genomes, which are known to contain large regions of distinct and widely different base composition (BERNARDI *et al.* 1989). Although the heterogeneity in regional base composition within Drosophila genomes is on the scale of that observed in warm blooded vertebrates, the prevalence or extent of this structuring along Drosophila chromosomes remains unclear.

In addition to our observation of regional heterogeneity in G + C content in the Drosophila genome, there is a tendency toward increasing G + C content with increasing distance from the centromere. Although the effect is not strong enough to lend statistical significance using the limited cytological distribution invoked in this study, it is interesting to compare this observation with results obtained for similar studies with other organisms. For human chromosomes, the most G + C-rich fraction of the genome maps to telomeric regions, and it has been suggested that the greatest concentration of genes is situated in the distal parts of all of human chromosomes (SACCONE, PESOLE and PREPARATA 1989). The data provided in this work cannot be used to directly address the relationship between gene density and base composition, though the genetic map of *D. melanogaster* reveals no systematic increase in the density of genes in the distal euchromatic regions of chromosomes (LINDSLEY and ZIMM 1991), and this may be an important difference in the organization of mammalian and drosophilid genomes. In addition, there is no

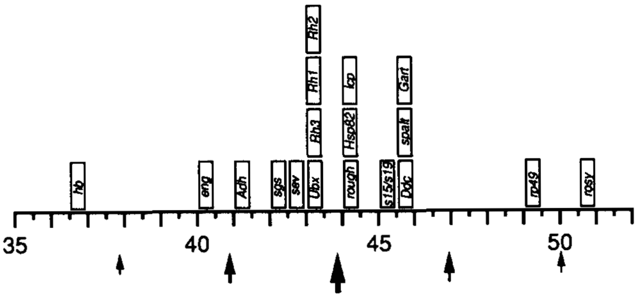

FIGURE 1.—Histogram of the distribution of G + C percentages for 18 YAC clones containing inserts of *D. melanogaster* genomic DNA. The large arrow indicates the mean, and smaller arrows are positioned at one and two SD from the mean. Abbreviations are the same as those used in Table 1, with the following additions: *hb* = *hunchback*, *eng* = *engrailed*, *sev* = *sevenless*. The YAC containing the *sgs* loci also contains *Sod*, and the YAC containing *Gart* also includes *pcp*.

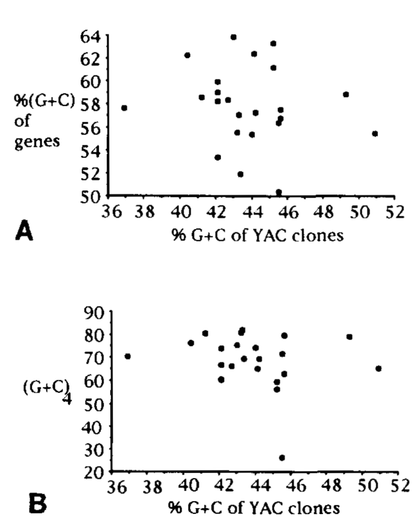

FIGURE 2.—(A) G + C percentage of YAC clones *vs.* the G + C content of coding sequences within the YACs ($r = -0.15$, $P = 0.5$, d.f. = 21). (B) Plot of the G + C percentage of YAC clones *vs.* the G + C percentage at fourfold degenerate nucleotide sites in genes within each YAC ($r = 0.162$, $P = 0.46$, d.f. = 21).

detectable relationship between chromosomal location and base composition at fourfold degenerate nucleotide sites (MORIYAMA and HARTL 1993). A more detailed and finer scale investigation of the distribution of compositional heterogeneity and gene density along a single Drosphila chromosome arm would be necessary to determine whether this observation reflects a true organizational property of the Drosphila genome.

Perhaps the most striking departure from mammalian systems of compositional heterogeneity is the lack of association of the base composition of chromosomal segments with base composition in coding regions or at synonymous sites within coding regions. For those YAC clones that include more than one sequenced gene, the nucleotide content at fourfold degenerate nucleotide sites varies over a range as great as 17.0%. Analyses complementary to those presented in this paper, including fine-scale comparisons of introns and adjacent exons from a large number of Drosophila genes, also fail to demonstrate any association between the base composition of genomic regions and the base composition of coding or noncoding DNA from the same chromosomal segments (SHIELDS *et al.* 1988; MORIYAMA and HARTL 1993). This is in contrast to what has been observed for mammalian genomes, where there is a very strong correlation between the base composition in DNA flanking genes, in introns and in synonymous sites of coding regions (IKEMURA 1985; BULMER 1987; AÏSSANI *et al.* 1991). This correlation has been cited as evidence for a mutational bias toward a specific G + C content in different chromosomal regions (BULMER 1987; HARDISON *et al.* 1991) and has also been used to argue that synonymous site evolution in mammals is strongly influenced by this mutational bias (WOLFE, SHARP and LI 1989).

Synonymous site evolution in Drosophila: Analysis of the relationship between regional base composition and the rate of nucleotide substitution in Drosophila implies a critical difference between the forces that generate variability in rates of DNA sequence evolution in Drosophila and in mammals. In contrast to the association between local base composition and evolutionary rate in mammals, there is no significant correlation between regional G + C content and the rate of synonymous site evolution in Drosophila. In mammals, the compartmentalization of the genome into large regions of homogeneous base composition has significant effects on rates and patterns of DNA sequence evolution. For example, comparison of the rates of synonymous site substitution among rodent genes shows that, for genes with a G + C content of more than 50%, the rate of synonymous site evolution decreases as the G + C content increases (WOLFE, SHARP and LI 1989). In Drosophila there is greater than twofold variation in synonymous site divergence among genes that have been compared between *D. melanogaster* and *D. pseudoobscura* (or *Drosophila subobscura*), and there is significant heterogeneity in base composition among the YAC clones that contain these genes. However, the lack of a statistically significant correlation between the G + C content and the rate of evolution of these genes suggests that factors other than the local base composition, or local mutational biases, are responsible for the variability in evolutionary rates.

If, unlike mammalian systems, the variability in evolutionary rates among Drosophila genes cannot be ascribed to differential mutational properties of distinct genomic regions—as broadly indicated by their base composition—what factors are associated with the observed patterns of synonymous site divergence? As suggested by SHIELDS *et al.* (1988), some form of selection acting at the level of synonymous codons

TABLE 2

Codon-usage bias and evolutionary rate of genes from *D. melanogaster* and *D. pseudoobscura*

	D. melanogaster		*D. pseudoobscura*			
Gene[a]	F_{op}	χ^2/L	F_{op}	χ^2/L	K_s (SE)[b]	K_a (SE)[c]
per	0.62	0.57	0.57	0.38	0.71 (0.06)	0.11 (0.009)
Gart	0.49	0.25	0.50	0.32	1.02 (0.08)	0.13 (0.009)
pcp	0.58	0.48	0.48	0.32	1.45 (0.51)	0.47 (0.028)
Adh	0.76	0.91	0.69	0.68	0.50 (0.08)	0.07 (0.013)
Hsp82	0.76	0.87	0.69	0.63	0.50 (0.08)	0.05 (0.009)
s15	0.53	0.31	0.46	0.56	0.70 (0.20)	0.37 (0.064)
s19	0.66	0.75	0.61	0.52	0.87 (0.20)	0.24 (0.034)
Rh4	0.58	0.35	0.60	0.40	0.73 (0.09)	0.06 (0.011)
bcd	0.44	0.22	0.45	0.24	0.99 (0.12)	0.13 (0.020)
rosy	0.51	0.26	0.61	0.49	0.93 (0.07)	0.11 (0.008)
Ubx	0.53	0.26	0.61	0.49	0.66 (0.10)	0.10 (0.015)
Rh2	0.50	0.26	0.65	0.57	0.77 (0.11)	0.11 (0.015)
Rh1	0.67	0.72	0.57	0.36	0.40 (0.06)	0.05 (0.010)
Rh3	0.57	0.55	0.56	0.40	0.62 (0.08)	0.10 (0.014)
rp49	0.77	1.00	0.63	0.58	0.61 (0.15)	0.05 (0.014)

[a] Gene designations and references as in Table 1, with the following additions: *per* = *period* (*D. melanogaster*, M30114: CITRI *et al.* 1987; *D. pseudoobscura*, X13878; COLOT, HALL and ROSBASH 1988); *Gart* and *pcp* (*D. pseudoobscura*, X06285: HENIKOFF and EGHTEDARZADEH 1987); *Adh* (*D. pseudoobscura*, Y00602; SCHAEFFER and AQUADRO 1987); *Hsp82* (*D. pseudoobscura*, X03812: BLACKMAN and MESELSON 1986); *s15* and *s19* (*D. suboboscura*, X53423: MARTINEZ-CRUZADO *et al.* 1988); *Rh4* = opsin protein (*D. melanogaster*, M17718, M17719: MONTELL *et al.* 1987; *D. pseudoobscura*, X65880, X65881: CARULLI and HARTL 1992); *bcd* = bicoid (*D. melanogaster*, X07870: BERLETH *et al.* 1988; *D. pseudoobscura*, entered manually by J.P.C. and D.E.K.: SEEGER and KAUFMAN, 1990); *rosy* (*D. pseudoobscura*, M33977: RILEY 1989); *Ubx* (*D. pseudoobscura*, X05179: WILDE and AKAM 1987); *Rh2, Rh1, Rh3* (*D. pseudoobscura*, X65878, X65677, X65879: CARULLI and HARTL, 1992); *rp49* (*D. subobscura*, M21333: AGUADÉ 1988).

[b] Synonymous site divergence and standard error (SE) of the divergence, determined by the method of LI, WU and LUO (1985).

[c] Nonsynonymous site divergence and standard error (SE) of the divergence.

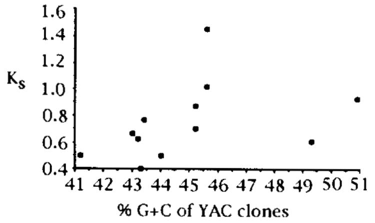

FIGURE 3.—G + C content of *D. melanogaster* YAC clones *vs.* the synonymous site divergence for genes compared between *melanogaster* group species and *obscura* group species ($r = 0.375$, $P = 0.24$, d.f. = 10).

may contribute to the diversity of evolutionary rates observed in Drosophila. The strong, negative correlation between codon-usage bias and synonymous site divergence suggests that highly biased genes in Drosophila are constrained in their evolution by incorporating only a limited subset of the potential codons. In bacteria, where the molecular basis for codon-usage bias has been examined, those codons that are used more frequently are those with the largest cellular tRNA pools (IKEMURA 1985). As a result, bacterial genes that are highly expressed have high codon-usage biases, and the reduced rate of synonymous site evolution in these genes may reflect selection for translational efficiency (SHARP and LI 1987). However, the molecular basis for codon-usage bias in a metazoan such as Drosophila is not easily explained, particularly since different tissues may have different tRNA pools,

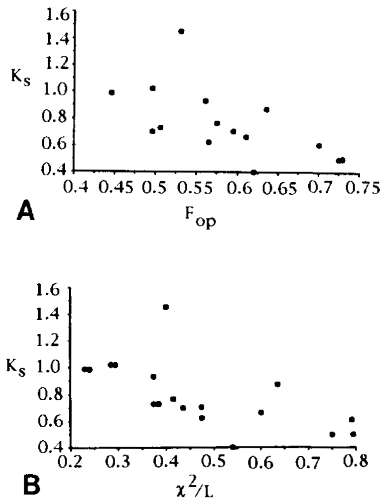

FIGURE 4.—Codon-usage bias, as measured by F_{op}, *vs.* synonymous site divergence (K_s) for genes compared between *melanogaster* and *obscura* group species ($r = -0.595$, $P = 0.02$, d.f. = 12). Codon-usage bias, as measured by χ^2/L, *vs.* synonymous site divergence for the same set of genes ($r = -0.605$, $P = 0.008$, d.f. = 12).

and levels of gene expression may differ among species or among populations of the same species (*e.g.*, FANG and BRENNAN 1992). Furthermore, codon-usage bias of homologous loci in *D. melanogaster* and *D. pseudoobscura* may differ significantly (CARULLI and HARTL 1992). Regardless, codon-usage bias does not appear to be a reflection of genome compartmentalization in Drosophila, because there is no association between codon-usage bias and regional base composition.

The nature of Drosophila compositional heterogeneity: The Drosophila genome is subdivided into large regions that differ from one another in base composition, and further analysis of compositional heterogeneity in the Drosophila genome may elucidate additional organizational features that are responsible for this pattern. The total genetic content of each of the chromosomal segments that we investigated is not known, so the observed heterogeneity in nucleotide content could be a function of the amount of coding DNA (which is relatively G + C rich) included in each of the yeast artificial chromosome clones, or the distribution of satellite DNA and other repetitive elements. Additional analysis of the distribution of coding sequences within the YAC clones will determine if coding capacity of the clones is responsible for the differences in G + C content. A hierarchical analysis of compositional heterogeneity at several levels of scale, including lambda or cosmid clones that span several hundred contiguous kilobases, may reveal that there are other patterns of heterogeneity that are not observed in comparisons of YAC clones.

Our data argue against a model of Drosophila genome organization and evolution that includes mutational biases that are purely a function of regional differences in base composition, but they do not suggest that rates of DNA sequence evolution are completely independent of genomic context. BEGUN and AQUADRO (1992) have recently demonstrated that the level of nucleotide polymorphism in *D. melanogaster* and its sibling species *Drosophila simulans* is strongly influenced by the variation in recombination rates in different chromosomal regions. The presence of chromosomal inversion polymorphisms within natural populations of Drosophila also affects the evolution of numerous genetic loci (AQUADRO *et al.* 1991). Therefore, variability in rates of synonymous site divergence among Drosophila genes incorporates the combined effects of differences in recombination rates, selection among alternative codons and perhaps other unknown factors. Our results, and those presented by SHIELDS *et al.* (1988) and by MORIYAMA and HARTL (1993), suggest fundamental differences in the forces that influence rates and patterns of molecular evolution in Drosophila and in mammals: variability in mutation rates, as a function of the compartmentalization of the genome into regions of distinct base composition and mutational properties, does not appear to be a dominant force contributing to the observed heterogeneity in rates of DNA sequence evolution in Drosophila.

We thank ETSUKO MORIYAMA for providing computer software and for critical reading of the manuscript, MARI BISERCIC for technical assistance in the screening of the YAC clones, and CHUNG-I WU for sharing computer programs. This research was supported by National Institutes of Health grants HG00750 to D.L.H., and GM-48407 to H.O. J.P.C. was supported by National Institutes of Health predoctoral training grant EY07108.

LITERATURE CITED

AGUADÉ, M., 1988 Nucleotide sequence comparison of the rp49 gene region between *Drosophila subobscura* and *D. melanogaster*. Mol. Biol. Evol. **5:** 433–441.

AÏSSANI, B., G. D'ONOFRIO, D. MOUCHIROUD, K. GARDINER, C. GAUTIER and G. BERNARDI, 1991 Compositional properties of human genes. J. Mol. Evol. **33:** 493–503.

AJIOKA, J. W., D. A. SMOLLER, R. W. JONES, J. P. CARULLI, A. E. VELLEK, D. GARZA, A. J. LINK, I. W. DUNCAN and D. L. HARTL, 1991 *Drosophila* genome project: one hit coverage in yeast artificial chromosomes. Chromosoma **100:** 495–509.

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proc. Natl. Acad. Sci. USA **88:** 305–309.

ASHBURNER, M., 1989 *Drosophila*. A Laboratory Manual. Cold Spring Harbor Laboratory. Cold Spring Harbor, N.Y.

BASLER, K., and E. HAFEN, 1988 Control of photoreceptor cell fate by the *sevenless* protein requires a functional tyrosine kinase domain. Cell **54:** 299–311.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

BERLETH, T., M. BURRI, G. THOMA, D. BOPP, S. RICHSTEIN, G. FRIGERIO, M. NOLL and C. NUSSLEIN-VOLLHARD, 1988 The role of localization of *bicoid* RNA in organizing the anterior pattern of the Drosophila embryo. EMBO J. **7:** 1749–1756.

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

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL and F. RODIER, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

BICKMORE, W., and A. T. SUMNER, 1989 Mammalian chromosome banding—an expression of genome organization. Trends Genet. **5:** 144–148.

BLACKMAN, R. K., and M. MESELSON, 1986 Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila hsp82* gene. J. Mol. Biol. **188:** 499–515.

BODMER, M., and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding for alcohol dhydrogenease in sibling species of *Drosophila*. Nature **309:** 425–430.

BOYLE, A. L., S. G. BALLARD and D. C. WARD, 1990 Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence *in situ* hybridization. Proc. Natl. Acad. Sci. USA **87:** 7757–7761.

BURKE, D. T., G. F. CARLE and M. V. OLSON, 1987 Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science **236:** 806–812.

BULMER, M., 1987 A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol. Biol. Evol. **4:** 395–405.

CARULLI, J. P., and D. L. HARTL, 1992 Variable rates of evolution among Drosophila opsin genes. Genetics **132:** 193–204.

CHU, G., D. VOLLRATH and R. W. DAVIS, 1986 Separation of large DNA molecules by contour-clamped homogeneous electric fields. Science **234:** 1582–1585.

CITRI, Y., H. V. COLOT, A. C. JACQUIER, Q. YU, J. C. HALL, D.

BALTIMORE and M. ROSBASH, 1987 A family of unusually spliced, biologically active transcripts encoded by a *Drosophila* clock gene. Nature **326:** 42–47.

COLOT, H. V., J. C. HALL and M. ROSBASH, 1988 Interspecific comparison of the *period* gene of *Drosophila* reveals large blocks of non-conserved coding DNA. EMBO J. **7:** 3929–3937.

COWMAN, A. F., C. S. ZUKER and G. M. RUBIN, 1986 A rhodopsin expressed in only one photoreceptor cell type of the *Drosophila* eye. Cell **44:** 705–710.

EVELETH, D. D., R. D. GIETZ, C. A. SPENCER, F. E. NARGANG, R. B. HODGETTS and J. L. MARSH, 1986 Sequence and structure of the *dopa decarboxylase* gene of *Drosophila*: evidence for novel RNA splicing variants. EMBO J. **5:** 2663–2672.

FANG, X., and M. BRENNAN, 1992 Multiple *cis*-acting sequences contribute to evolved regulatory variation for Drosophila *Adh* genes. Genetics **131:** 333–343.

FEINBERG, A. P., and B. VOGELSTEIN, 1983 A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:** 6–13.

FILIPSKI, 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.

FREI, E., R. SCHUH, S. BAUMGARTNER, M. BURRI, M. NOLL, G. JUERGENS, E. SEIFERT, U. NAUBER and H. JAECKLE, 1988 Molecular characterization of *spalt*, a homeotic gene required for head and tail development in the *Drosophila* embryo. EMBO J. **7:** 197–204.

GARFINKEL, M. D., R. E. PRUITT and E. M. MEYROWITZ, 1983 DNA sequences, gene regulation, and modular protein evolution in the *Drosophila* 68C glue gene cluster. J. Mol. Biol. **168:** 765–789.

GARZA, D., J. W. AJIOKA, D. T. BURKE and D. L. HARTL, 1989 Mapping the *Drosophila* genome with yeast artificial chromosomes. Science **246:** 641–646.

GOLDMAN, M. A., G. P. HOLMQUIST, M. C. GRAY, L. A. CASTON and A. NAG, 1984 Replication timing of genes and middle repetetive sequences. Science **234:** 686–692.

HARDISON, R., D. KRANE, D. VANDENBERGH, J.-F. CHENG, J. MANSBERGER, J. TADDIE, S. SCHWARTZ, X. HUANG and W. MILLER, 1991 Sequence and comparative analysis of the rabbit α-like globin gene cluster reveals a rapid mode of evolution in a G + C-rich region of mammalian genomes. J. Mol. Biol. **222:** 233–249.

HARTL, D. L., 1992 Genome map of *Drosophila melanogaster* based on yeast artificial chromosomes, pp. 39–69 in *Genome Analysis*, Vol. 4, edited by K. E. DAVIES and S. M. TILGHMAN. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

HENIKOFF, S., and M. K. EGHTEDARZADEH, 1987 Conserved arrangement of nested genes at the Drosophila *Gart* locus. Genetics **117:** 711–725.

HENIKOFF, S., M. A. KEENE, K. FECHTEL and J. W. FRISTROM, 1986 Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell **44:** 33–42.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–34.

IKEMURA, T., and S.-I. AOTA, 1988 Global variation in G + C content along vertebrate genome DNA: possible correlation with chromosome band structures. J. Mol. Biol. **203:** 1–13.

IKEMURA, T., K.-W. WADA and S.-I. AOTA, 1990 Giant G + C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. Genomics **8:** 207–216.

KEITH, T. P., M. A. RILEY, M. KREITMAN, R. C. LEWONTIN, D. CURTIS and G. CHAMBERS, 1987 Sequence of the structural gene for xanthine dehydrogenas (*rosy* locus) in *Drosophila melanogaster*. Genetics **116:** 67–73.

KRANE, D. E., D. L. HARTL and H. OCHMAN, 1991 Rapid determination of nucleotide content and its application to the study of genome structure. Nucleic Acids Res. **19:** 5181–5185.

KWIATOWSKI, J., M. PATEL and F. J. AYALA, 1989 *Drosophila melanogaster* Cu-Zn superoxide mutase gene sequence. Nucleic Acids. Res. **17:** 1264.

LEE, C. S., D. CURTIS, M. MCCARRON, C. LOVE, M. GRAY, W. BENDER and A. CHOVNICK, 1987 Mutations affecting expression of the *rosy* locus in *Drosophila melanogaster*. Genetics **116:** 55–66.

LEWONTIN, R. C., 1985 Population genetics. Annu. Rev. Genet. **19:** 81–102.

LEWONTIN, R. C., 1991 Twenty-five years ago in genetics. Electrophoresis in the development of evolutionary genetics. Genetics **128:** 657–662.

LI, W.-H., C.-I. WU and C.-C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

LINDSLEY, D. L., and G. G. ZIMM, 1992 *The Genome of Drosophila melanogaster*. Academic Press, San Diego, Calif.

MARTIN, C. H., and E. M. MEYEROWITZ, 1986 Characterization of the boundaries between rapidly and slowly evolving genomic regions of *Drosophila*. Proc. Natl. Acad. Sci. USA **83:** 8654–8658.

MARTINEZ-CRUZADO, J., C. SWIMMER, M. G. FENERJIAN and F. C. KAFATOS, 1988 Evolution of the autosomal chorion locus in *Drosophila*. I. General organization of the locus and sequence comparisons of genes *s15* and *s19* in evolutionarily distant species. Genetics **119:** 663–677.

MONTELL, C., K. JONES, C. S. ZUKER and G. M. RUBIN, 1987 A second opsin gene expressed in the ultraviolet sensitive R7 cells of Drosophila melanogaster. J. Neurosci. **7:** 1558–1566.

MORIYAMA, E. M., and D. L. HARTL, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics (in press).

O'TOUSA, J. E., E. BAEHR, R. L. MARTIN, J. HIRSCH, W. L. PAK and M. L. APPLEBURY, 1985 The *Drosophila ninaE* gene encodes an opsin. Cell **40:** 839–850.

POOLE, S. J., L. M. KAUVAR, D. DREES and T. KORNBERG, 1985 The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. Cell **40:** 37–43.

RILEY, M. A., 1989 Nucleotide sequence of the *Xdh* region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol. Biol. Evol. **4:** 638–650.

SACCONE, C., G. PESOLE and G. PREPARATA, 1989 DNA microenvironments and the molecular clock. J. Mol. Evol. **29:** 407–411.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS and H. A. EHRLICH, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487–491.

SCHAEFFER, S. W., and C. F. AQUADRO, 1987 Nucleotide sequence of the *Adh* region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics **117:** 61–73.

SCHILDKRAUT, C. L., and J. J MAIO, 1969 Fractions of HeLa DNA differing in their content of guanine and cytosine. J. Mol. Biol. **46:** 305–312.

SEEGER, M., and T. C. KAUFMAN, 1990 Molecular analysis of the *bicoid* region from *Drosophila pseudoobscura*: identification of conserved domains within coding and noncoding regions. EMBO J. **9:** 2977–2987.

SHARP, P. M., and W.-H. LI, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

SHARP, P. M., and W.-H. LI, 1989 On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28:** 398–402.

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.

SNYDER, M., M. HUNKAPILLER, D. YUEN, D. SILVERT, J. FRISTROM and N. DAVIDSON, 1982 Cuticle protein genes of *Drosophila*: structure, organization and evolution of four clustered genes. Cell **29:** 1027–1040.

SOKAL, R. R. and F. J. ROHLF, 1969 *Biometry*. W. H. Freeman, San Francisco.

SORSA, V., 1988 *Chromosome Maps of Drosophila*. CRC Press, Boca Raton, Fla.

TAUTZ, D., R. LEHMANN, H. SCHNUERCH, R. SCHUH, E. SEIFERT, A., KIENLIN, K. JONES and H. JACKLE, 1987 Finger protein of novel structure encoded by ***hunchback***, a second member of the gap class of *Drosophila* segmentation genes. Nature **327:** 383–389.

THIERY, J.-P., G. MACAYA and G. BERNARDI, 1976 An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol. **108:** 219–235.

WILDE, C. D., and K. AKAM, 1987 Conserved sequence elements in the 5′ region of the *ultrabithorax* transcription unit. EMBO J. **6:** 1393–1401.

WONG, Y-C., J. PUSTELL, N. SPOEREL and F. C. KAFATOS, 1985 Coding and potential regulatory sequences of a cluster of chorion genes in *Drosophila melanogaster*. Chromosoma **92:** 124–135.

WOLFE, K. H., P. M. SHARP and W.-H. LI, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

ZUKER, C. S., A. F. COWMAN and G. M. RUBIN, 1985 Isolation and structure of a rhodopsin gene from *Drosophila*. Cell **40:** 852–858.

ZUKER, C. S., C. MONTELL, K. JONES, T. LAVERTY and G. M. RUBIN, 1987 A rhodopsin gene expressed in photorecteptor cell R7 of the *Drosophila* eye: homologies with other signal transducing molecules. J. Neurosci. **7:** 1550–1557.

Communicating editor: A. G. CLARK