Potentials and Limitations of Histone Repeat Sequences for Phylogenetic Reconstruction of *Sophophora*

*Angela M. Baldo,**[1] *Donald H. Les,*† *and Linda D. Strausbaugh**

*Department of Molecular and Cell Biology and †Department of Ecology and Evolutionary Biology, University of Connecticut

Simplified DNA sequence acquisition has provided many new data sets that are useful for phylogenetic reconstruction, including single- and multiple-copy nuclear and organellar genes. Although transcribed regions receive much attention, nontranscribed regions have recently been added to the repertoire of sequences suitable for phylogenetic studies, especially for closely related taxa. We evaluated the efficacy of a small portion of the histone repeat for phylogenetic reconstruction among *Drosophila* species. Histone repeats in invertebrates offer distinct advantages similar to those of widely used ribosomal repeats. First, the units are tandemly repeated and undergo concerted evolution. Second, histone repeats include both highly conserved coding and variable intergenic regions. This composition facilitates application of ''universal'' primers spanning potentially informative sites. We examined a small region of the histone repeat, including the intergenic spacer segments of coding regions from the divergently transcribed H2A and H2B histone genes. The spacer (about 230 bp) exists as a mosaic with highly conserved functional motifs interspersed with rapidly diverging regions; the former aid in alignment of the spacer. There are no ambiguities in alignment of coding regions. Coding and noncoding regions were analyzed together and separately for phylogenetic information. Parsimony, distance, and maximum-likelihood methods successfully retrieve the corroborated phylogeny for the taxa examined. This study demonstrates the resolving power of a small histone region which may now be added to the growing collection of phylogenetically useful DNA sequences.

Introduction

Nuclear ribosomal genes are tandemly arranged in many organisms and are used extensively for phylogenetic reconstruction (Maley and Marshall 1998). Such sequences offer technical advantages over single-copy genes, given their high concentration in the genome. Moreover, the repeats are mosaics of regions of low and high divergence, permitting a wide range of phylogenetic applications. The architectures and sequences of such repeats allow the use of universal PCR primers in conserved regions to span variable regions containing phylogenetically informative sites (Schlötterer et al. 1994). In this context, an important feature of tandem arrays is that repeats are essentially homogeneous within individuals and within a species, exhibiting a pattern called concerted evolution (Linares, Bowen, and Dover 1994).

Histone genes exist in tandem, multicopy arrays in many invertebrates. They also share other relevant features with ribosomal repeats, such as high copy number and a mosaic structure, and offer an additional feature in translated coding regions. The similarity between histone and rDNA repeats suggests that the former may also be well suited for phylogenetic reconstruction. We considered the DNA sequence of the spacer region of the divergently transcribed gene pair H2A-H2B an attractive candidate for several reasons. This spacer has a small, conserved length of 225–300 bp in *Drosophila* species and is flanked by highly conserved coding regions, an ideal arrangement for the placement of universal PCR primers and internal sequencing primers. Equally important, extensive functional information exists for the spacer (Sommer 1996), thus facilitating alignment and evolutionary interpretation of these regions.

A prerequisite for the use of tandemly repeated genes in phylogenetic applications is a high level of identity between repeats due to concerted evolution. The earliest indication that fly histone repeats undergo concerted evolution was genomic blot analysis of *Drosophila melanogaster*. Digestion with single enzymes that recognize one site per repeat generates 5-kb-unit- length fragments (Lifton et al. 1978; Strausbaugh and Weinberg 1979). Genomic blots of other *Sophophora* (including *Drosophila mauritiana, Drosophila simulans, Drosophila yakuba, Drosophila pseudoobscura,* and *Drosophila paulistorum*) also show the diagnostic uniformity of concerted evolution (Strausbaugh and Weinberg 1979; Coen, Strachan, and Dover 1982). Representatives from radiations outside *Sophophora,* such as *Drosophila hydei, Drosophila virilis,* and *Drosophila hawaiiensis,* also appear to have multiple uniform copies (Domier et al. 1986; Fitch, Strausbaugh, and Barrett 1990). Concerted evolution of histone genes has been further demonstrated by the mapping and sequencing of cloned repeats in *D. melanogaster*. Matsuo and Yamazaki (1989*a*) observed uniform geometry of restriction endonuclease recognition sites in cloned histone repeats and also found a low proportion of polymorphism (0.0159) in the 1,200 bp containing the H3 gene and adjacent spacers. Goldberg (1979) and Matsuo and Yamazaki (1989*b*) separately cloned and sequenced entire repeats from different strains, revealing nearly identical sequences. Although *Drosophila* histone loci are not without variation (Strausbaugh and Weinberg 1979; Col-

[1] Present address: Department of Plant Breeding, Cornell University.

Abbreviation: DSE, downstream element.

Key words: *Drosophila,* systematics, molecular evolution, promoter.

Address for correspondence and reprints: Linda D. Strausbaugh, Box U-131, University of Connecticut, Storrs, Connecticut 06269-2131. E-mail: strausba@uconnvm.uconn.edu.

Mol. Biol. Evol. 16(11):1511–1520. 1999

by and Williams 1993), the available evidence suggests a level of concerted evolution comparable to that in rDNA.

A well-corroborated phylogeny is necessary to suitably evaluate the efficacy of histone repeat sequences for phylogenetic inference. The *Sophophora* radiation includes four species groups (Melanogaster, Obscura, Willistoni, and Saltans) whose phylogeny has been assessed thoroughly (Anderson, Carew, and Powell 1993; Kwiatowski et al. 1994; Caccone et al. 1996). Species subgroups within each of these species groups have been defined. Relationships within the subgroups Melanogaster, Obscura, and Willistoni have been identified and confirmed by a variety of data, ranging from chromosome morphology to mitochondrial sequence analysis data (Cohn, Thompson, and Moore 1984; Tsakas and Tsacas 1984; Felger and Pinsker 1986; Maruyana and Hartl 1991; Peixoto et al. 1992; Barrio, Latorre, and Moya 1994; Eickbush and Eickbush 1995; Ritchie and Gleason 1995; Clark, Leicht, and Muse 1996; Okuyama et al. 1996).

Here, we examine the phylogenetic resolution provided by a small segment of the histone repeat for members of the well-corroborated *Sophophora.* This study provides a framework to assess future exploitation of histone genes for phylogenetic inference.

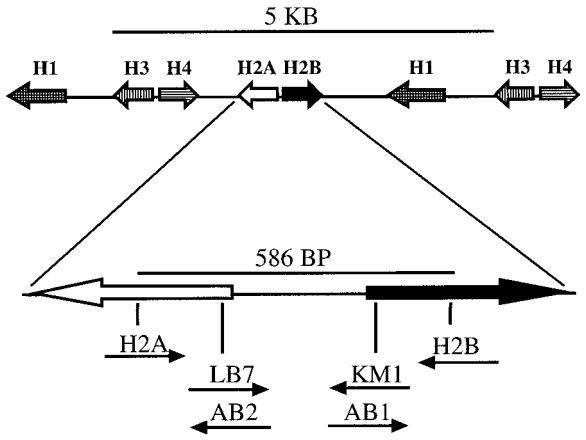

FIG. 1.—The 5-kb histone gene repeat in *Sophophora* consists of one gene for each of the four nucleosomal core proteins (H2A, H2B, H3, and H4) and the linker protein (H1). Amplification primers H2A and H2B are positioned 586 bp apart in conserved areas of the coding region in each gene. Internal primers LB7 and KM1, also in conserved regions, are used for sequencing the intergenic spacer. Internal primers AB1 and AB2 provide additional sequence for the flanking coding regions. Genes are designated above corresponding arrows.

Materials and Methods

Drosophila Stocks

The following strains were obtained from the National Species Stock Center, Bowling Green, Ohio: *D. melanogaster* (14021-0231.0), *D. simulans* (14021-0251.0), *D. mauritiana* (14021-0214.0), *D. yakuba* (14021-0261.0), *Drosophila kikkawai* (14028-0561.0), *Drosophila seguyi* (14028-0671.1), *D. pseudoobscura* (14011-0121.0), *Drosophila persimilis* (14011-0111.0), *D. paulistorum* (14030-0771.0), *Drosophila equinoxialis* (14030-0741.0), *Drosophila tropicalis* (14030-0801.0), *Drosophila pavlovskiana,* and *Drosophila insularis.* Some alternative isolates for the Willistoni species group were provided by Dr. Lee Ehrman (State University of New York at Purchase) and Mr. Stephen Daniels (University of Connecticut). Male and female representatives from selected stocks were examined (external genitalia, sex combs) to verify species assignments.

Preparation of Genomic DNA

Multiple extractions were carried out for each species isolate using a protocol modified from Daniels and Strausbaugh (1986). DNA extractions were checked by electrophoresis in a 1% agarose TBE gel and visualized by ethidium bromide staining. DNA concentrations were determined using a Hoefer DyNA Quant 200 fluorimeter.

Amplification and Sequencing of Histone H2A-H2B Intergenic Spacer and Flanking Coding Regions

Histone repeat features and the primers used in their study are depicted in figure 1. Two universal primers (H2A: 5′- GCAGCATTGCCAGCCAACT-3′; H2B: 5′-CTGTTCATTATGCTCATCGCCTT-3′) were used to amplify fragments from genomic DNA. A combination of three forward (H2A; LB7: 5′-CCTTTCCCTTCACTTTGCCACC-3′; and AB1: 5′-AGTGGAAAGGCAGCCAAGAA-3′) and three reverse (H2B; KM1: 5′-TTCTTGGCTGCCTTTCCACT-3′; and AB2: 5′-GGTGGCAAAGTGAAGGGAAAGG-3′) universal fluorescently tagged primers were used in automated sequencing of the PCR product. Standard primers were used for manual sequencing. All primers were synthesized by the Macromolecular Characterization Facility of the Biotechnology Center, University of Connecticut.

Histone fragments were amplified from 100 ηg of genomic DNA with 0.5 ηg of H2A and H2B primers in 50 μl PCR reactions using a Perkin Elmer Cetus (PEC) Gene Amp PCR reagent kit. Amplification products were separated from PCR components, primers, and unincorporated nucleic acids with a 1000 NML spin-column (Millipore) or with a Bio 101 Gene Clean kit (La Jolla, Calif.) per manufacturers' instructions. Purified products were resuspended in 200 μl of deionized water.

Sequencing reactions were carried out directly on 5–10 μl of the resuspended PCR product in a model 2400 thermocycler (PEC) with an Amplitaq cycle sequencing kit (PEC) per manufacturer's instructions, or manually with a U.S. Biochemicals (USB) Sequenase kit, following standard procedures. Each base was confirmed by a minimum of two reactions in each direction. PCR and sequencing artifacts were identified and corrected by comparison of sequences generated by independently amplified samples.

Phylogenetic Analysis

Sequences were initially aligned by computer using CLUSTAL W 1.6 (Higgins, Thompson, and Gibson 1996). A unique artificial sequence was introduced at each end of the sequence to enhance the program's abil-

ity to recognize and align flanking sequences, despite gaps in the spacer. The alignment was then manually adjusted. As a check, two scientists independently made adjustments. Alignments are available from the EMBL Nucleotide Sequence Database under accession number DS33312.

Phylogenetic analyses were performed with test version 4.0.0d57 of PAUP* with the permission of author David L. Swofford. All characters in parsimony analyses were treated as unordered (Fitch) and were optimized by ACCTRAN (multiple states = uncertainty). Furthest addition sequence was used, and the MULPARS option was in effect. All searches were performed using the branch-and-bound algorithm. Bootstraps were obtained from 500 replicates using a heuristic search. Decay indices were calculated by observing collapse of nodes in strict-consensus trees obtained by sequentially filtering all sets of trees 1–10 steps longer than the shortest tree.

Distance analyses employed minimum evolution as the objective function. All substitutions were included. The heuristic search was used with starting trees obtained by neighbor joining and tree bisection-reconnection (TBR) branch swapping. We evaluated uncorrected (*p*) distances and Jukes-Cantor, Kimura three-parameter, and GTR models both with and without rate heterogeneity which was assumed to follow a gamma distribution. The proportion of invariant sites was assumed to be 0. Shape parameters were estimated using maximum likelihood (see below). Bootstraps were obtained from 500 replicates using heuristic search.

Maximum-likelihood analyses used empirical base frequencies and a two-parameter model (HKY-85) for unequal frequencies. Transition/transversion ratios were estimated, and a molecular clock was not enforced. Analyses were conducted with and without equal rates. Among-site rate variation was assumed to follow a gamma distribution. The proportions of invariable sites and shape parameters were estimated with six rate categories represented by means. Starting branch lengths were determined using the Rogers-Swofford approximation. Heuristic searches were conducted with TBR branch swapping and MULPARS in effect. Bootstraps were obtained from 500 replicates using quick addition.

Analyses of all 13 taxa were performed using three data partitions: coding regions only (270 bp), spacer regions only (209–267 bp), and combined coding and spacer regions (586 bp).

Results

Sequence Data

Spacer lengths from 209 to 267 bp and flanking coding regions (270 bp) were sequenced for 13 species. Two separate geographical isolates were sequenced for *D. paulistorum.* Replicate extracts of genomic DNA were prepared from each strain using 200 individuals. Each DNA sample was amplified a minimum of two times, and amplification products were directly sequenced several times. This experimental design enabled PCR and sequencing errors to be traced and eliminated. To detect and correct for this phenomenon, each amplification product was sequenced and compared with other sequences from the same isolate. PCR errors were detected as sequence variation unique to one template. The sequence for each species was confirmed a minimum of two times on each strand with nested and overlapping primers (fig. 1); the majority of sequences were confirmed three times in each direction. Eighty-seven percent of the ambiguities were resolved by a majority consensus from the multiple sequences. In unresolved cases, the base was designated as ambiguous. Unresolved ambiguities ranged from 0% to 2% depending on the species and may represent compound technical errors or polymorphic sites. In these unresolved cases, PCR artifacts occurred at a rate of 0.5% based on repeated sequencing of individual PCR amplifications. Sequences are available from the EMBL Nucleotide Sequence Database under the following accession numbers: *D. mauritiana*: AJ224806; *D. simulans*: AJ224807; *D. melanogaster*: AJ224808; *D. yakuba*: AJ224809; *D. seguyi*: AJ224810; *D. kikkawai*: AJ224811; *D. pseudoobscura*: AJ224812; *D. persimilis*: AJ224813; *D. paulistorum*: AJ224814; *D. pavlovskiana*: AJ224815; *D. tropicalis*: AJ224816; *D. equinoxialis*: AJ224817; *D. insularis*: AJ224818.

Alignment

Alignment of conserved coding regions was unambiguous. As is the case with rDNA, computer alignment of the noncoding region is complicated by the interspersion of conserved and nonconserved regions. CLUSTAL W aligns most conserved regions well, but more divergent regions were aligned best among closely related species and very poorly among more distant relatives, yielding a high genetic distance between distantly related species. We then manually adjusted the alignment to reflect known promoter and other regulatory motifs; we considered regions with the same putative transcriptional and translational functions and similar sequences as homologous. Nonconserved regions diverged not only in sequence, but also in length; the manual adjustment was accomplished by adding a net 29 bases in gaps, taking particular care not to disrupt known promoter motifs. Manual adjustment of the CLUSTAL alignment produced a more conservative alignment between distant taxa and resulted in a lower amount of divergence between every taxon pair. One of the independent adjustments was accomplished without extensive familiarity with either *Drosophila* taxonomy or eukaryotic promoters. Despite the latter ''blind'' control, the resulting alignments were in excellent agreement and consistently reflected conservation of identifiable promoter motifs. Regions of discrepancy tended to involve shifts of single bases that had no effect on the resultant phylogenies. It should be emphasized here that the same topology was recovered regardless of the slightly different alignment variants used.

Figure 2 illustrates one probable alignment that clearly illustrates the mosaic nature of the spacer and emphasizes some of the regulatory motifs used in alignment. Sequences in a regulatory region that are required

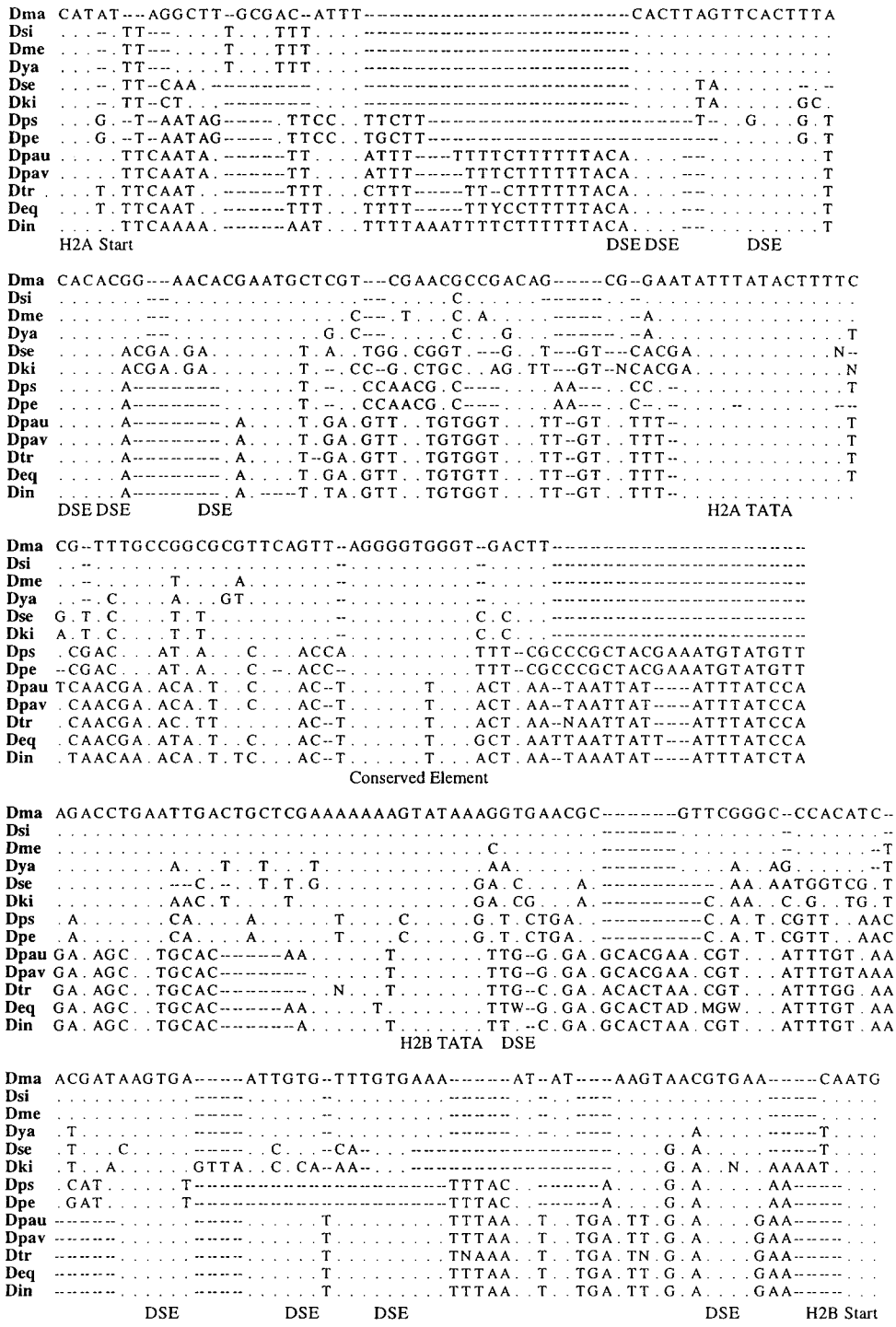

FIG. 2.—Alignment of H2A-H2B spacer in the 13 taxa of this study. H2A and H2B start codons are labeled, as are putative promoter elements including TATA boxes, downstream elements (DSEs), and a central, highly conserved element.

for transcriptional or translational functions will be conserved compared with surrounding areas and may correspond to known motifs. Several such examples are evident in the H2A-H2B spacer. A TATA box can be observed upstream of each gene (5′-GTATAATA-3′ 99–73 bp upstream of H2A, and 5′-GTATAAA-3′ 63–87 bp upstream of H2B). The divergent nucleotides in the spacer largely follow clade-specific patterns. Two to seven short, clustered motifs (5′-GTG-3′) are present upstream of both genes between the start codons and the TATA boxes. This repeating GTG motif, also called a downstream element (DSE), has been associated with TATA-less genes in *D. melanogaster* (Arkhipova 1995). The most striking feature is the conserved element AGGGGT(T)GGGT in the center of the spacer. The only divergence present in this region is the addition of a T

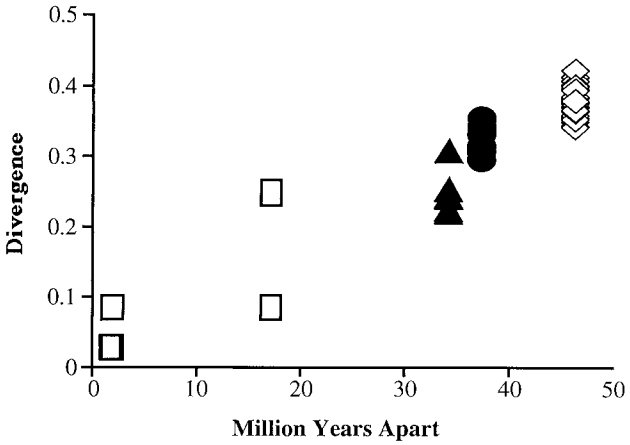

FIG. 3.—Divergence versus millions of years apart in the H2A-H2B intergenic spacer. Divergence times reflect published estimates. □ ,within species subgroups; ▲, between species subgroups; ●, between Melanogaster and Obscura groups; ♦, between Old World *Sophophora* and New World *Sophophora*.There is a linear correlation in the *Sophophora* between levels of nucleic acid changes and time since evolutionary divergence. Cricket graph generates a linear fit ($y = 1.4995e^{-2} + 7.7433e^{-3x}$; $R^2 = 0.894$).

in the New World *Sophophora*. We have evidence that this motif is involved in transcriptional regulation (Sommer 1996).

While no alignment of this nature is 100% accurate, and some subjectivity is introduced by our adjustment, the data set we present is robust. We have also examined several different alignments that represent only slight variations (in ambiguous regions) of the one presented and find consistent support for the clades regardless of alignments. This indicates that the regions which present the most difficult alignment problems contributed little phylogenetic information. The coding regions, which present no problems with alignment, recover precisely the same topology. While the coding region alone resolves species groups, adding the spacer provides more resolution within species groups. This indicates that our alignment of the spacer region effectively recovers the phylogenetic signal from the data. Obviously, the closer the phylogenetic relationships between the species, the easier the alignment of the spacer, making this region of the histone repeat particularly attractive for resolving closely related groups.

Sequence Divergence and Phylogenetic Inference

Sequences useful for phylogenetic inference should contain levels of divergence that correlate with the accepted phylogeny. The amount of divergence in the spacer (summarized in fig. 3) correlates well with estimated times of divergence, approximately 0.0075 bases per Myr. The frequency of transitions relative to transversions shows little trend at low levels of divergence but drops as divergence increases as expected when transversions slowly replace transitions in multiple hits (DeSalle et al. 1987). In accordance with the potential for multiple hits, the nucleotide composition is different between the Old and New World *Sophophora*. Pooled data from *Sophophora* generate a χ^2 $P < 0.005$ for homogeneity. When a similar test is used within each group, $P > 0.975$, not departing from homogeneity. This nucleotide composition is unlikely to confound phylogenetic inference, as about half of the difference is due to a scattered, asymmetrical enrichment for T on one strand in the Willistoni group (10% more than for the Old World *Sophophora*) at equal expense of G and C. The remaining difference is due to an insert in members of the Willistoni group that is slightly upstream of the H2A start, containing 9/10 bases that are T (fig. 2).

An appropriate test of a data set for phylogenetic inference is the validation of its performance on an accepted phylogeny. Such a test determines the resolving power of the sequence at various taxonomic levels. Each of three data sets (spacer, coding, and combined) was challenged to retrieve the corroborated phylogeny by a variety of methods: maximum parsimony, distance, and maximum likelihood. PAUP* was used to generate estimates of a gamma distribution for among-site rate variation for each data set. Using six rate categories and estimated values of invariant sites, there was less among-site rate variation in the spacer (alpha = 19.623032), and predictably more (alpha = 0.449533) in the coding region, where most divergent nucleotides occur in silent positions.

The coding region alone is insufficient to retrieve all details of the corroborated topology with any of the methods, although major features of the topologies were consistent with all of the methods used. The coding region sequenced (270 bp, 90 codons) provides too little signal to resolve relationships below the level of species groups (fig. 4 and table 1). The spacer (316 bp), however, retrieves the accepted phylogeny with a variety of methods, including maximum parsimony, distance (uncorrected *p*), and maximum likelihood (equal rates). Use of a gamma rate distribution and/or the Jukes-Cantor, Kimura three- parameter, or GTR model impairs distance and maximum-likelihood methods from retrieving some details of the corroborated phylogeny. The combined data set (586 bp) proved the most robust. It generated the corroborated topology with gamma-corrected maximum likelihood and gamma-corrected and equal-rates Jukes-Cantor, Kimura three-parameter, and GTR models and all of the methods mentioned above that were successful with the spacer. We were unable to completely resolve the Willistoni species group with any methods.

Discussion

The major objective of this work was to ascertain whether a small portion of the histone repeat comprising the H2A-H2B intergenic spacer (209–267 bp) contains sufficient information for phylogenetic reconstruction. Among the most widely used genome sequences for generating evolutionary hypotheses are repeated sequences, long noted for homogeneity within species while being substantially different between even closely related species (Collins and Paskewitz 1996). The taxonomically widespread advantages of abundance, homogeneity, and conserved orientation in histone genes

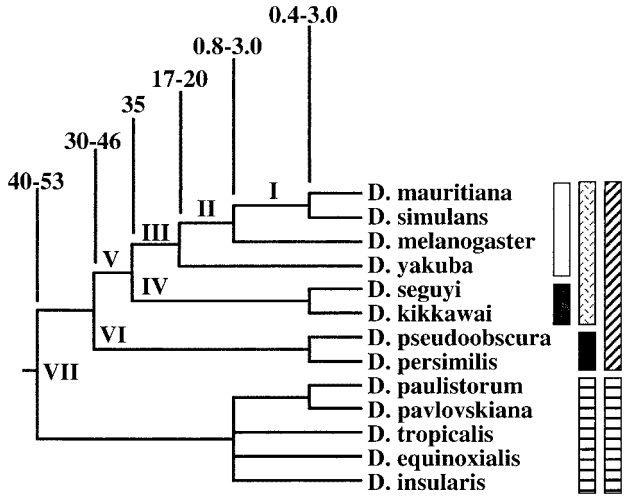

FIG. 4.—Trees retrieved match the corroborated phylogeny. The corroborated phylogeny of *Sophophora*, based on strict consensus from morphology, behavior, chromosome morphology, DNA-DNA hybridization, and nuclear and mitochondrial genes (summarized in Caccone et al. [1996], and from other sources, including Peixoto et al. [1992]; Anderson, Carew, and Powell [1993]; Kwiatowski et al. [1994]; Eickbush and Eickbush [1995]; Ritchie and Gleason [1995]; Clark, Leicht, and Muse [1996]; and Okuyama et al. [1996]). Divergence time estimates (MYA) are indicated. Taxa are designated as follows: diagonals, Old World *Sophophora*; horizontals, New World *Sophophora*; stippled, Melanogaster species group; solid, Obscura species group; horizontals, Willistoni species group; open, Melanogaster species subgroup; shaded, Montium species subgroup. Branches are labeled with Roman numerals to correspond with bootstrap values and decay indices reported in table 1.

suggest that this multigene family has potential for phylogenetic exploitation. This work solidifies the position of the tandemly repeated histone genes in the growing repertoire of nuclear multigene families that show promise for molecular systematics. As is the case for any repetitive sequence that is relatively new to molecular systematic uses, it is important to demonstrate that the forces of concerted evolution render the region appropriate for such applications.

This is not the first time that histone sequences have been employed in phylogenetic applications, but it is the first use of the H2A-H2B spacer in such a context. Amino acid sequences of the slowly evolving nucleosomal histone proteins have long been used as molecular measures of very distant relationships (Thatcher and Gorovsky 1994; del Gaudio et al. 1998). More recently, the structure and organization of histone genes and their sequences have also been used (del Gaudio et al. 1998). The spacer between H3 and H4 has been used to infer relationships among ciliates (Brunk, Kahn, and Sadler 1990). This study demonstrates the resolving power of a very small intergenic region that lends itself well to amplification with a ''universal'' PCR primer set. We have focused on invertebrate taxa separated by less than 60 Myr and have demonstrated that a very small amount of sequence information (about 230 bases) can generate a tree in agreement with the corroborated phylogeny. One limitation to the spacer region is that alignment by computer programs may be confounded by its mosaic structure, and manual adjustments may be necessary. This intergenic spacer is particularly valuable for closely related species groups in which alignment is more easily accomplished. In comparison, twice as many rDNA sequence data were shown to provide a similar amount of resolution by Pélandakis, Higgins, and Solignac (1991), although this may have been due in part to the wide range of taxa considered in their study.

The amount of pairwise divergence present in the H2A-H2B spacer (0.018–0.259 within species groups and 0.351–0.419 between species groups) is comparable with that found in other genes from the same taxa. In general, the H2A-H2B spacer evolves more slowly than do the Amylase genes (Okuyama et al. 1996) and the rDNA internal transcribed spacer (Schlötterer et al. 1994). The rate is similar to that of mitochondrial NADH dehydrogenase 2 and cytochrome oxidase 1 genes (Satta and Takahata 1990) and to the intron present in myosin alkali light chain 1 (Clark, Leicht, and Muse 1996), zeste, and per (Hey and Kliman 1993). The spacer evolves slightly faster than yolk protein 2 (Hey and Kliman 1993). The H2A-H2B spacer thus evolves at a rate similar to those of many genes that have already proven informative. In the case of this system, resolution is limited to species diverged less than 60 Myr, which translates to roughly $\leq$40% divergence in the intergenic spacer. The limitation is due to the rearrangement of conserved putative regulatory motifs in other radiations and the associated difficulty in alignment. Even within 60 Myr, it is very helpful to have sequences from several representatives within each species group in order to identify conserved regions, which can then be aligned between species groups.

The knowledge that the H2A and H2B genes in many invertebrates are paired and transcribed divergently promises a potential for widespread application of this system. Histone repeat clusters have been characterized in many organisms (fig. 5), including Echinodermata, Insecta, Crustacea, Annelida, and Cnidaria (del Gaudio et al. 1998). Histone repeats typically consist of one gene for each of the four nucleosomal core proteins (H2A, H2B, H3, and H4) and their associated spacers. Many phylogenetic groups also carry a gene for the linker protein H1. The organization of nucleosomal core histone genes is fairly conserved across invertebrates. All of the previously mentioned groups (with the exception of Echinodermata) have repeating units with H2A and H2B arranged as one divergently transcribed gene pair and H3 and H4 arranged as a second divergently transcribed gene pair. Elements of the repeat unit evolve at different rates. The amino acid sequences of core histones H3 and H4 evolve most slowly, with H2A and H2B evolving 10 times as fast and H1 evolving even more rapidly (Thatcher and Gorovsky 1994). The DNA sequences of the intergenic spacers evolve the most rapidly of all (Kremer and Hennig 1990; Fitch and Strausbaugh 1993). Additional sequence information from the histone repeat may be sufficient to resolve relationships in the Willistoni group.

There are many evolutionary radiations, especially in invertebrates, that currently lack adequate resolution for phylogenetic relationships. Many insect vectors, for

Table 1
Bootstrap Values for the Nodes in Figure 4 Trees as Generated by PAUP*

		I	II	III	IV	V	VI	VII
Parsimony								
Spacer (316 bp)	BS	89	96	85	100	99	100	100
	(DI)	(2)	(6)	(4)	(9)	(9)	(>10)	(>10)
Coding (270 bp)	BS	26	43	99	95	84	100	100
	(DI)	(0)[a]	(0)[a]	(6)	(4)	(2)	(7)	(10)
Combined (586 bp)	BS	90	97	99	100	100	100	100
	(DI)	(2)	(6)	(10)	(>10)	(>10)	(>10)	(>10)
Distance (uncorrected *p*)								
Spacer (316 bp)		85	96	100	100	95	100	100
Coding (270 bp)		25	[b]	100	99	72	100	100
Combined (586 bp)		99	94	100	100	95	100	100
Distance (equal rates) combined data sets (586 bp)								
Jukes-Cantor		94	94	100	100	94	100	100
Kimura three-parameter		95	95	100	100	95	100	100
GTR		93	93	100	100	94	100	100
Distance (gamma-corrected) combined data sets (586 bp)								
Jukes-Cantor		55	76	97	100	90	100	100
Kimura three-parameter		[b]	58	94	100	90	100	100
GTR		[b]	62	72	100	91	100	100
Maximum likelihood (equal rates)								
Spacer (316 bp)		89	84	71	100	95	98	100
Coding (270 bp)		30[b]	33[b]	97	96	84	100	100
Combined (586 bp)		89	87	95	100	100	100	100
Maximum likelihood (gamma-corrected)								
Spacer (316 bp)		78	70	53	99	89	90	100
Coding (270 bp)		57	57	92	89	53	100	98
Combined (586 bp)		72	72	77	99	93	99	100

NOTE.—Each bootstrap was accomplished with 500 replicates and the heuristic search algorithm. Decay indices are supplied for parsimony trees. BS = bootstrap value; (DI) = decay index.

[a] Resolved differently in two trees (collapsed in consensus).

[b] Unresolved in the search.

example, are members of morphologically identical sibling species, of which only some participate in the transmission of disease (Hill and Crampton 1994). While rDNA has been used for bioidentification of *Anopheles* species (Collins and Paskewitz 1996), histone gene spacers could be used to support and potentially enhance available resolution. Useful phylogenetic hypotheses for more distantly related taxa are likely with other data from the histone repeat. Regions of the histone repeat that contain the core histone genes and their intergenic spacers are small in invertebrates (less than 5 kb), suggesting that it is entirely feasible to mine the wealth of information that is useful at different phylogenetic levels with a relatively small amount of sequence acquisition.

For many phylogenetic problems, it is useful to include more than one type of data (Maley and Marshall 1998). It is similarly valuable to study sequences evolving under different evolutionary pressures. Each new gene that is demonstrated to corroborate known phylogenies at a given taxonomic level is a valuable addition. Histone genes are subject to their own unique constraints. Unlike nuclear single-copy genes, they are kept in homogeneous abundance by concerted evolution; yet, unlike rDNA, they are translated into proteins. In addition, understanding the evolutionary dynamics of molecules is fundamental to developing efficient approaches to their use in a phylogenetic context (Hillis and Moritz 1990). The histone proteins are exceedingly well understood at a functional molecular level, and the spacer region we employed in this study has been extensively dissected in our lab for functional components (Sommer 1996). Given the expense of collection and confirmation of DNA sequence data, the availability of a small region with high resolving power is extremely valuable. In this regard, the H2A-H2B intergenic spacer offers a particularly compelling system.

Acknowledgments

We thank Dr. Kent Holsinger and Dr. Chris Simon of the University of Connecticut at Storrs for helpful suggestions and stimulating discussions, and Dr. Rob DeSalle and Dr. Jeff Powell for providing valuable insights in the initial stages. We also acknowledge Dr. Laurine Bow, whose doctoral research formed the foundation for this work, and Valerie Schwaroch at the American Museum of Natural History, New York, for assistance in dissection. We thank the undergraduate and high school students who assisted in various stages of this work, especially Karen Morse, Kristin Borodezt, Alex Kentsis, Rebecca Osthus, and Debbie Cebrik. We also thank Dr. David Swofford for allowing us to use a

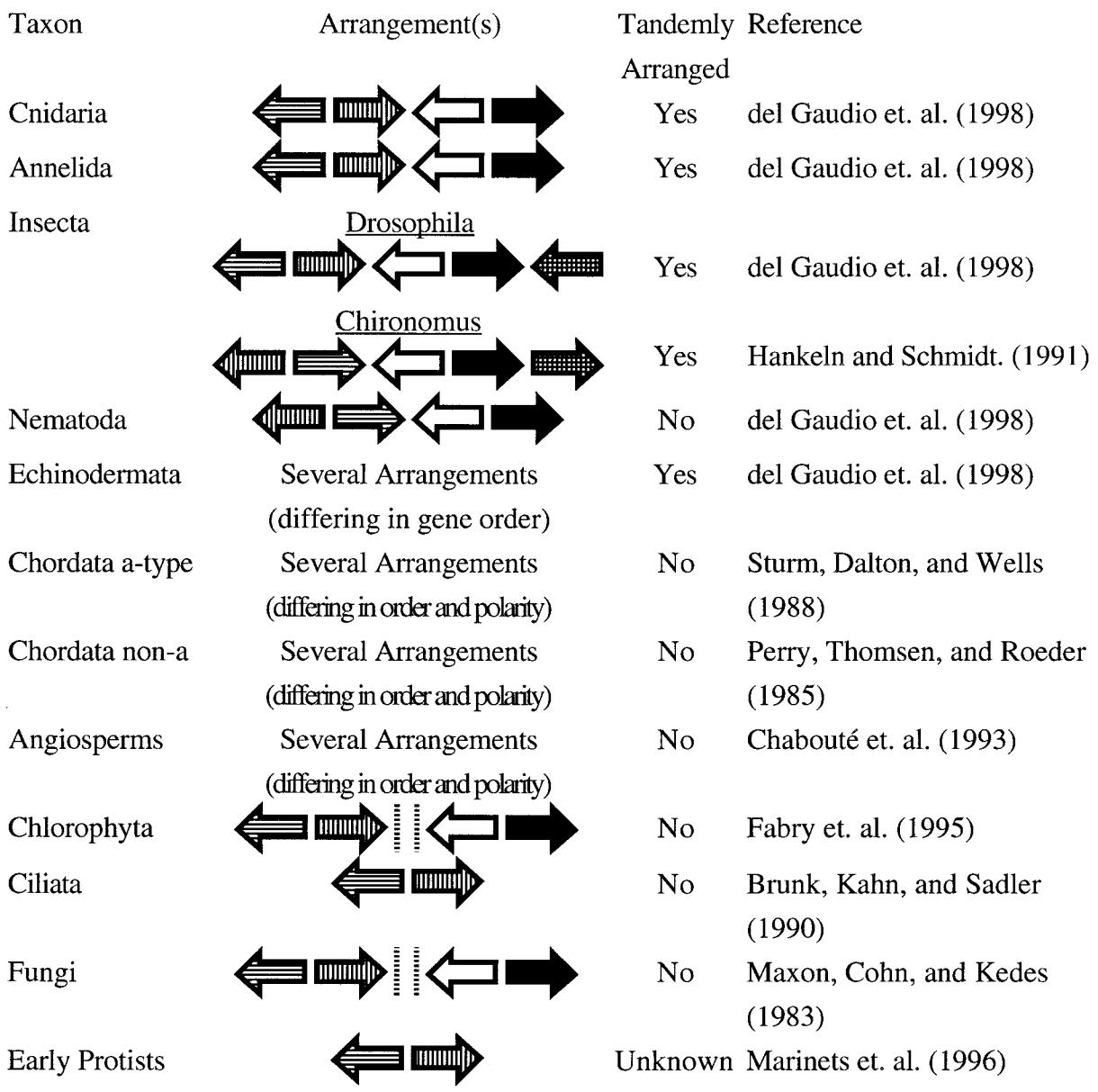

FIG. 5.—Features of histone gene structure in various organisms. Many organisms contain H2A/H2B and H3/H4 divergent gene pairs. Genes are designated as in Figure 1. Noncontiguous clusters are separated by two dotted lines.

test version of PAUP*. We acknowledge the support of the Alfred P. Sloan Foundation, the National Science Foundation (BSR-9009938), and the University of Connecticut Research Foundation in sponsoring this research.

LITERATURE CITED

ANDERSON, C. L., E. A. CAREW, and J. R. POWELL. 1993. Evolution of the ADH locus in the *Drosophila-willistoni* group—the loss of an intron, and shift in codon usage. Mol. Biol. Evol. **10**:605–618.

ARKHIPOVA, I. R. 1995. Promoter elements in *Drosophila melanogaster*. Genetics **139**:1359–1369.

BARRIO, E., A. LATORRE, and A. MOYA. 1994. Phylogeny of the *Drosophila obscura* species group deduced from mitochondrial DNA sequences. J. Mol. Evol. **39**:478–488.

BRUNK, C. F., R. W. KAHN, and L. A. SADLER. 1990. Phylogenetic relationships among *Tetrahymena* species determined using the polymerase chain reaction. J. Mol. Evol. **30**:290–297.

CACCONE, A., N. MORIYAMA, J. M. GLEASON, L. NIGRO, and J. R. POWELL. 1996. A molecular phylogeny for the *Drosophila melanogaster* subgroup and the problem of polymorphism data. Mol. Biol. Evol. **13**:1224–1232.

CHABOUTE, M. E., N. CHAUBET, C. GIGOT, and G. PHILIPPS. 1993. Histones and histone genes in higher plants: structure and genomic organization. Biochimie **75**:523–531.

CLARK, A. G., B. G. LEICHT, and S. V. MUSE. 1996. Length variation and secondary structure of introns in the *Mlc1* gene in six species of *Drosophila*. Mol. Biol. Evol. **13**:471–482.

COEN, E., T. STRACHAN, and G. DOVER. 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila.* J. Mol. Biol. **158**:17–35.

COHN, V. H., M. A. THOMPSON, and G. P. MOORE. 1984. Nucleotide sequence comparison of the Adh gene in three *Drosophilids.* J. Mol. Evol. **20**:31–37.

COLBY, C., and S. M. WILLIAMS. 1993. The distribution and spreading of rare variants in the histone multigene family of *Drosophila melanogaster*. Genetics **135**:127–133.

COLLINS, F. H., and S. M. PASKEWITZ. 1996. A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic *Anopheles* species. Insect Mol. Biol. **5**:1–9.

DANIELS, S. B., and L. D. STRAUSBAUGH. 1986. The distribution of P-element sequences in *Drosophila*: the *willistoni* and *saltans* species groups. J. Mol. Evol. **23**:138–148.

DEL GAUDIO, R., N. POTENZA, P. STEFANONI, M. L. CHIUSANO, and G. GERACI. 1998. Organization and nucleotide sequence of the cluster of five histone genes in the polichaete worm Chaetopterus variopedatus: first record of a H1 histone gene in the phylum Annelida. J. Mol. Evol. **46**:64–73.

DESALLE, R., T. FREEDMAN, E. M. PRAGER, and A. C. WILSON. 1987. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila.* J. Mol. Evol. **26**:157–164.

DOMIER, L. L., J. J. RIVARD, L. M. SABATINI, and M. BLUMFIELD. 1986. *Drosophila virilis* histone gene clusters lacking H1 coding segments. J. Mol. Evol. **23**:149–158.

EICKBUSH, D. G., and T. H. EICKBUSH. 1995. Vertical transmission of the retrotransposable elements *R1* and *R2* during the evolution of the *Drosophila melanogaster* species subgroup. Genetics **139**:671–684.

FABRY, S., K. MULLER, A. LINDAUER, P. BUM PARK, T. CORNELIUS, and R. SCHMITT. 1995. The organization structure and regulatory elements of *Chlamydomonas* histone genes reveal features linking plant and animal genes. Curr. Genet. **28**:333–345.

FELGER, I., and W. PINSKER. 1986. Histone gene transposition in the phylogeny of the *Drosophila obscura* group. Z. Zool. Syst. Evol. **25**:127–140.

FITCH, D. H. A., and L. D. STRAUSBAUGH. 1993. Low codon bias and high rates of synonymous substitution in *Drosophila hydei* and *Drosophila melanogaster* histone genes. Mol. Biol. Evol. **10**:397–413.

FITCH, D. H. A., L. D. STRAUSBAUGH, and V. BARRETT. 1990. On the origins of tandemly repeated genes: does histone gene copy number in *Drosophila* reflect chromosomal location? Chromosoma **99**:118–124.

GOLDBERG, M. 1979. Sequence analysis of *Drosophila* histone genes. Ph.D. thesis, Stanford University, Stanford, Calif.

HANKELN, T., and E. R. SCHMIDT. 1991. The organization, localization and nucleotide sequence of the histone genes of the midge *Chironomus thummi.* Chromosoma **101**:25–31.

HEY, J., and R. M. KLIMAN. 1993. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. **10**:804–822.

HIGGINS, D. G., J. D. THOMPSON, and T. J. GIBSON. 1996. Using Clustal for multiple sequence alignments. Meth. Enzymol. **266**:383–402.

HILL, S. M., and J. M. CRAMPTON. 1994. DNA-based methods for the identification of insect vectors. Ann. Trop. Med. Parasitol. **88**:227–250.

HILLIS, D. M., and C. MORITZ, eds. 1990. Molecular systematics. 1st edition. Sinauer, Sunderland, Mass.

KREMER, H., and W. HENNIG. 1990. Isolation and characterization of a *Drosophila hydei* histone DNA repeat unit. Nucleic Acids Res. **18**:1573–1580.

KWIATOWSKI, J., D. SKARECKY, K. BAILEY, and F. J. AYALA. 1994. Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the Cu,Zn *Sod* Gene. J. Mol. Evol. **38**:443–454.

LIFTON, R. P., M. L. GOLDBERG, R. W. KARP, and D. S. HOGNESS. 1978. The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications. Cold Spring Harb. Symp. Quant. Biol. **42**:1047–1051.

LINARES, A. R., T. BOWEN, and G. A. DOVER. 1994. Aspects of nonrandom turnover involved in the concerted evolution of intergenic spacers within the ribosomal DNA of *Drosophila melanogaster*. J. Mol. Evol. **39**:151–159.

MALEY, L. E., and C. R. MARSHALL. 1998. The coming of age of molecular systematics. Science **279**:505–506.

MARINETS, A., M. MULLER, P. J. JOHNSON, J. KULDA, O. SCHEINER, G. WIEDERMANN, and M. DUCHÊNE. 1996. The sequence and organization of the core histone H3 and H4 genes in the early branching amitochondriate protist *Trichomonas vaginalis.* J. Mol. Evol. **43**:563–571.

MARUYANA, K., and D. L. HARTL. 1991. Evolution of the transposable element mariner in *Drosophila* species. Genetics **128**:319–329.

MATSUO, Y., and T. YAMAZAKI. 1989*a.* Nucleotide variation and divergence in the histone multigene family in *Drosophila melanogaster*. Genetics **122**:87–97.

———. 1989*b.* tRNA derived insertion element in histone gene repeating unit of *Drosophila melanogaster*. Nucleic Acids Res. **17**:225–238.

MAXSON, R., R. COHN, and L. KEDES. 1983. Expression and organization of histone genes. Ann. Rev. Genet. **17**:239–277.

OKUYAMA, E., H. SHIBATA, H. TACHIDA, and T. YAMAZAKI. 1996. Molecular evolution of the 5′-flanking regions of the duplicated *Amy* genes in *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **13**:574–583.

PEIXOTO, A. A., R. COSTA, D. A. WHEELER, J. C. HALL, and C. P. KRYIACOU. 1992. Evolution of the threonine-glycine repeat region of the *period* gene in the *melanogaster* species subgroup of *Drosophila.* J. Mol. Evol. **35**:411–419.

PÉLANDAKIS, M., D. G. HIGGINS, and M. SOLIGNAC. 1991. Molecular phylogeny of the subgenus *Sophophora* of *Drosophila* derived from the large subunit of ribosomal RNA sequences. Genetica **84**:87–94.

PERRY, M., G. H. THOMSEN, and R. G. ROEDER. 1985. Genomic organization and nucleotide sequence of two distinct histone gene clusters from *Xenopus laevis*: identification of novel conserved upstream sequence elements. J. Mol. Biol. **185**:479–499.

RITCHIE, M. G., and J. M. GLEASON. 1995. Rapid evolution of courtship song pattern in *Drosophila willistoni* sibling species. J. Evol. Biol. **8**:463–497.

SATTA, Y., and N. TAKAHATA. 1990. Evolution of *Drosophila* mitochondrial DNA and the history of the *melanogaster* subgroup. Proc. Natl. Acad. Sci. USA **87**:9558–9562.

SCHLÖTTERER, C., M.-T. HAUSER, A. VON HAESELER, and D. TAUTZ. 1994. Comparative evolutionary analysis of rDNA ITS regions in *Drosophila.* Mol. Biol. Evol. **11**:513–522.

SOMMER, M. T. 1996. Replication-dependent regulation of the histone H2A gene via promoter—transcription factor interactions in *Drosophila melanogaster*. Ph.D. thesis, University of Connecticut, Storrs.

STRAUSBAUGH, L. D., and E. S. WEINBERG, eds. 1979. Heterogeneity in histone gene organization in *Drosophila*. Vol. 14. Academic Press, New York.

STURM, R. A., S. DALTON, and J. R. E. WELLS. 1988. Conservation of histone H2A/H2B intergene regions: a role for the H2B specific element in divergent transcription. Nucleic Acids Res. **16**:8571–8586.

THATCHER, T. H., and M. A. GOROVSKY. 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res. **22**:174–179.

TSAKAS, S. C., and L. TSACAS. 1984. A phenetic tree of eighteen species of the *melanogaster* group of *Drosophila* using allozyme data as compared with classifications based on other criteria. Genetica **64**:139–144.

THOMAS H. EICKBUSH, reviewing editor

Accepted July 27, 1999