Birth-and-Death Evolution with Strong Purifying Selection in the Histone H1 Multigene Family and the Origin of *orphon* H1 Genes

José M. Eirín-López, Ana M. González-Tizón, Andrés Martínez, and Josefina Méndez

Departamento de Biología Celular y Molecular, Universidade da Coruña, Campus de A Zapateira, A Coruña, Spain

Histones are small basic nuclear proteins with critical structural and functional roles in eukaryotic genomes. The H1 multigene family constitutes a very interesting histone class gathering the greatest number of isoforms, with many different arrangements in the genome, including clustered and solitary genes, and showing replication-dependent (RD) or replication-independent (RI) expression patterns. The evolution of H1 histones has been classically explained by concerted evolution through a rapid process of interlocus recombination or gene conversion. Given such intriguing features, we have analyzed the long-term evolutionary pattern of the H1 multigene family through the evaluation of the relative importance of gene conversion, point mutation, and selection in generating and maintaining the different H1 subtypes. We have found the presence of an extensive silent nucleotide divergence, both within and between species, which is always significantly greater than the nonsilent variation, indicating that purifying selection is the major factor maintaining H1 protein homogeneity. The results obtained from phylogenetic analysis reveal that different H1 subtypes are no more closely related within than between species, as they cluster by type in the topologies, and that both RD and RI H1 variants follow the same evolutionary pattern. These findings suggest that H1 histones have not been subject to any significant effect of interlocus recombination or concerted evolution. However, the diversification of the H1 isoforms seems to be enhanced primarily by mutation and selection, where genes are subject to birth-and-death evolution with strong purifying selection at the protein level. This model is able to explain not only the generation and diversification of RD H1 isoforms but also the origin and long-term persistence of *orphon* RI H1 subtypes in the genome, something that is still unclear, assuming concerted evolution.

Introduction

Histones are small basic nuclear proteins, ubiquitous in all eukaryotic species, that are involved in the packaging of DNA and also in the regulation of gene expression. There are five major classes, which can be classified into two groups according to their functional and structural features: core histones (H2A, H2B, H3 multigene families, and H4 gene family) and linker histones (H1 multigene family). With the exception of the H4 histone, for which variants have not been described, histones can be classified on the basis of their genomic organization and expression patterns as replication-dependent (RD), actively expressed during the S-phase of the cell cycle, and as replication-independent (RI), expressed at low levels but continuously throughout the cell cycle (Isenberg 1979; Maxson, Cohn, and Kedes 1983; Doenecke et al. 1997). Also stage-specific and tissue-specific histones can be defined, which are specifically expressed during early embryogenesis and in particular cell types, respectively (Hentschel and Birnstiel 1981; D'Andrea et al. 1985; Ohsumi and Katagiri 1991).

The H1 histone multigene family encodes linker proteins, which bind to the linker DNA in the chromatin fiber constituting the chromatosomal structure. There are multiple H1 isoforms, which have been best characterized in mammals whose complement consists of five somatic subtypes (H1.1 to H1.5), a tissue-specific subtype (H1t), a replacement subtype (H1°), and an oocyte-specific subtype (H1oo) (Albig et al. 1997; Wang et al. 1997; Tanaka et al. 2001). In nonmammalian species, there is a second differentiation-specific subtype (H5) related to H1° and expressed only in avian and amphibian nucleated erythrocytes (reviewed by Khochbin and Wolffe [1994]) and also another oocyte-specific H1 histone known as B4 or H1M (maternal) (Dimitrov et al. 1993). In invertebrates, the lower complexity determines the presence of fewer H1 isoforms, which are only defined by punctual changes of amino acid residues at specific positions. In the case of plants, many H1 genes possess intervening sequences (introns), the presence of polyadenylation signals in the mRNA is the rule rather than the exception, and there are several stress-inducible H1 subtypes (reviewed by Chabouté et al. [1993]).

Although the H1 multigene family is the fastest-evolving class among histones, H1 proteins are still highly conserved proteins and concerted evolution has been invoked to explain its evolution (Kedes 1979; Hentschel and Birnstiel 1981; Coen, Strachan, and Dover 1982; Ohta 1983; Hankeln and Schmidt 1993; Schienman, Lozovskaya, and Strausbaugh 1998). However, many multigene families do not fit the predictions made by the concerted-evolution hypothesis, and sequences of gene members are more closely related between than within species. To account for these observations, Nei and Hughes (1992) first proposed a new evolutionary model that they named the "birth-and-death" model of evolution. In this model, new genes are created by repeated gene duplication, and some of the duplicate genes are maintained in the genome for a long time, whereas others are deleted or become nonfunctional. Protein homogeneity is maintained by the effect of the strong purifying selection, and, consequently, DNA sequences of different members can be very different, both within and between species (Nei and Hughes 1992; Nei, Gu, and Sitnikova 1997; Nei, Rogozin, and Piontkivska 2000). This model has been reported as the primary mode of evolution for several multigene families, such as the major histocompatibility complex (MHC) (Nei and Hughes 1992; Gu and Nei 1999), immunoglobulin (Ota and Nei 1994), antibacterial ribonuclease genes (Zhang, Dyer, and Rosenberg 2000), nematode chemoreceptor gene families

Key words: birth-and-death evolution, purifying selection, concerted evolution, histone H1, *orphon* genes.

E-mail: che@udc.es.

Mol. Biol. Evol. 21(10):1992–2003. 2004
doi:10.1093/molbev/msh213
Advance Access publication July 14, 2004

(Robertson 2000), ubiquitins (Nei, Rogozin, and Piontkivska 2000), T-cell receptor (Su and Nei 2001), histone 3 multigene family (Rooney, Piontkivska, and Nei 2002), histone 4 gene family (Piontkivska, Rooney, and Nei 2002), elapid snake venom three-finger toxins (Fry et al. 2003), plant MADS-box genes (Nam et al. 2004), and heat-shock 70 proteins from nematodes (Nikolaidis and Nei 2004). Although concerted evolution and birth-and-death evolution are conceptually different, they may not be distinguishable if the rate of concerted evolution is assumed to be very slow. In this work, we define concerted evolution as a rapid process of interlocus recombination or gene conversion so that even related species have different sets of homogeneous member genes (Dover 1982).

The purpose of this work is to provide a deeper insight into the long-term evolutionary pattern of the H1 multigene family through the evaluation of the relative importance of gene conversion, point mutation, and selection using the above criteria. In this sense, the presence of such independent RI H1 variants represents an invaluable tool used to test whether concerted evolution or birth-and-death evolution guides the long-term evolution of the H1 multigene family. The present contribution completes the molecular evolutionary characterization of the H1 histone multigene family and its *orphon* variants discussed in two previous reports by Eirín-López et al. (2002, 2004).

Materials and Methods

We have included in our analysis all the nonredundant nucleotide H1 sequences listed in the NHGRI/NCBI Histone Sequence Database (Sullivan et al. 2002) as of December 2003 (see table in Supplementary Material online). Sequences retrieved were subsequently corrected for errors in accession numbers and nomenclature. There are no less than 12 different nomenclatures for the H1 subtypes, but to reach the broadest audience possible and a certain homogeneity with our previous works, we have used Doenecke laboratory's numeric nomenclature (Albig, Meergans, and Doenecke 1997) in the present work. The alignment of nucleotide sequences was constructed on the basis of the translated amino acid sequences using the programs BIOEDIT (Hall 1999) and ClustalX (Thompson et al. 1997). This alignment consisted of a set of 146 sequences belonging to 55 different species, showing 1,362 nucleotide sites, excluding the start and stop codons. Additionally, the corresponding protein alignment consisted of a set of 144 sequences (because of the presence of two pseudogenes) showing 456 amino acid positions. Alignments were visually inspected for errors in both cases. All the molecular evolutionary analyses in this work were conducted using the computer program MEGA version 2.1 (Kumar et al. 2001). The extent of nucleotide and amino acid sequence divergence was estimated by means of the uncorrected differences (*p*-distance) because this distance is known to give better results than more complicated distances when the number of sequences is large and the number of positions used is relatively small, because of its smaller variance (Nei and Kumar 2000). The numbers of synonymous (p_S) and nonsynonymous (p_N) nucleotide differences per site were computed using the modified Nei-Gojobori method (Zhang, Rosenberg, and Nei 1998), providing in both cases the transition/transversion ratio (R). Both amino acid and nucleotide distances were estimated using the pairwise-deletion option, and standard errors were calculated by the bootstrap method (1,000 replicates). The presence of positive selection was analyzed by testing the null hypothesis that H_0: $p_S = p_N$, being the alternative that H_1: $p_S > p_N$. The average p_S and p_N values and also their variances were compared using the codon based Z-test for selection (Nei and Kumar 2000). The Z-statistic and the probability that the null hypothesis is rejected were obtained, being this probability indicated as $**P$ ($P < 0.001$) and $*P$ ($P < 0.05$).

Phylogenetic trees were reconstructed using the neighbor-joining (NJ) tree-building method (Saitou and Nei 1987). The reliability of the resulting topologies were tested by the bootstrap method (Felsestein 1985) and by the interior-branch test (Rzhetsky and Nei 1992; Sitnikova 1996), which produced the bootstrap probability (BP) and confidence probability (CP) values, respectively, for each interior branch in the tree. Because the bootstrap method is known to be conservative, BP $> 80\%$ was interpreted as high statistical support for interior branches in the tree, CP $= 95\%$ was otherwise considered statistically significant (Sitnikova, Rzhetsky, and Nei 1995). We rooted phylogenetic trees using the H1 gene of the protist *Entamoeba histolytica*, as it represents one of the most primitive eukaryotes for which an H1-related protein has been characterized (Kasinsky et al. 2001).

The GenBank database and complete genome databases (chicken, human, mouse, rat, *Drosophila*, nematode, sea urchin, *Arabidopsis*, corn, tomato, and wheat) were screened for the presence of H1 pseudogenes using the Blast tool (Altschul et al. 1990). The presence of truncated or incomplete H1 sequences, indels in the conserved protein central domain, as well as the absence or interruption of the major H1 5′ promoter elements were viewed as pseudogenization features used to define putative H1 pseudogenes.

Results

H1 Protein Evolution

The phylogenetic tree for H1 proteins was reconstructed from 144 amino acid sequences of 55 species belonging to different eukaryotic kingdoms (fig. 1). The different taxonomic groups are well defined in the topology on basis of their H1 proteins. Although plant and invertebrate H1 proteins still do not show clear differences among subtypes, it is possible to discriminate among H1s more closely related between than within species in the cases of the H1-I protein from *Glyptotendipes barbipes* and *G. salinus*, the H1e protein from *Chironomus tentans* and *C. pallidivittatus*, and the stress-inducible H1 variants from the plants *Lycopersicon esculentum* and *Lycopersicon chilense*.

In vertebrates, there is clear functional differentiation among the isoforms, being evident the presence of a monophyletic origin for all the RD H1 proteins but the human H1.X histone. In the case of mammals, all the proteins were encoded by orthologous genes in the phylogeny cluster by type and not by species, where the

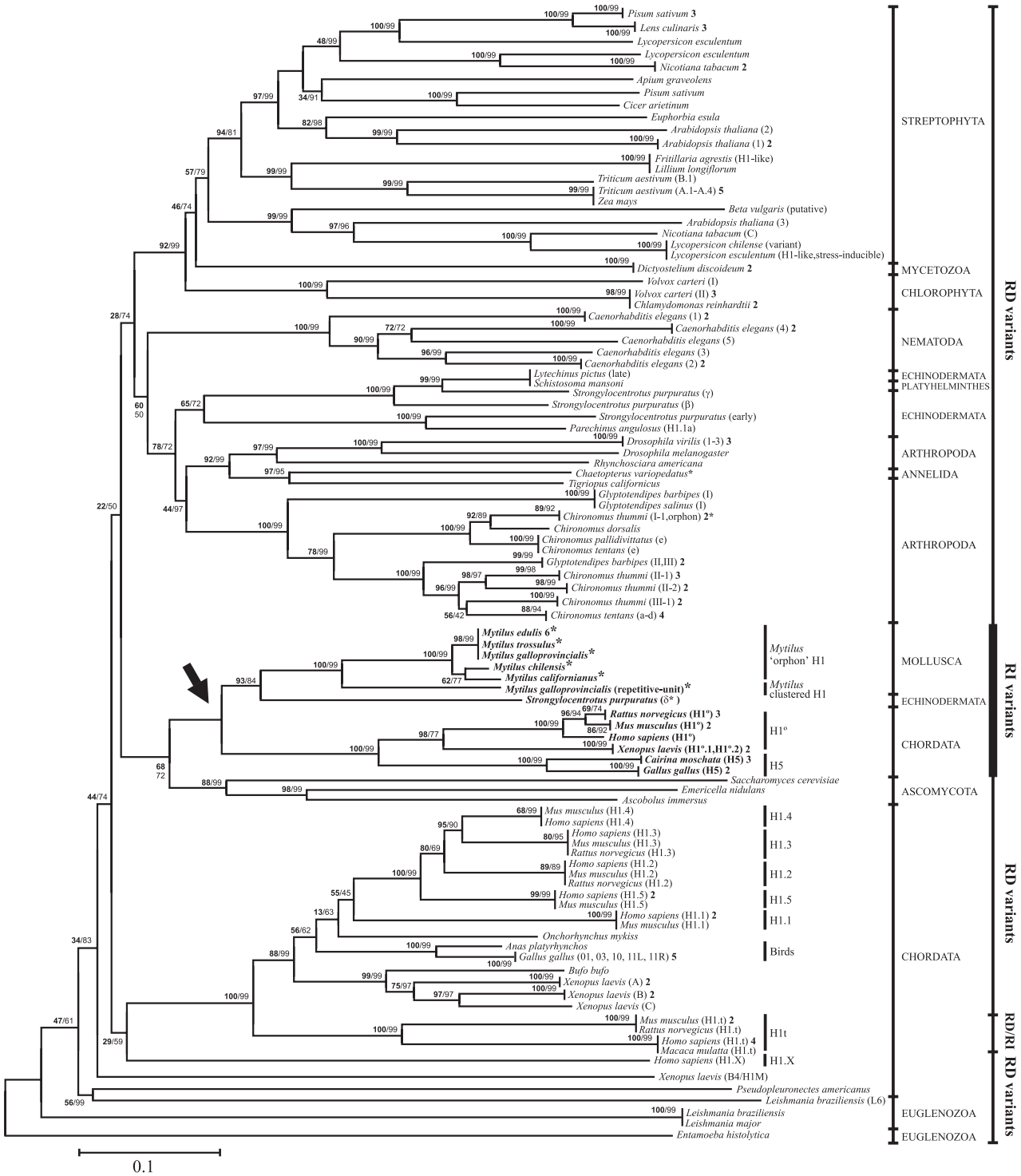

FIG. 1.—Phylogenetic relationships among H1 proteins from all the organisms analyzed using uncorrected *p*-distances. The numbers for interior branches represent BP values (boldface), followed by CP interior-branch test values (normal) based on 1,000 replications, and are only shown when a value is greater than 50%. Numbers in parentheses near species indicate the H1 subtype and in boldface, the number of sequences analyzed for each species. Possible invertebrate RI H1 genes are marked by asterisks (*). The black arrow indicates the origin of the monophyletic group gathering the RI H1 variants. Taxonomic groups, vertebrate subtypes, as well as expression patterns are indicated in the right margin of the tree.

groups of H1.1 to H1.5, H1.X, and H1t subtypes are well defined and statistically supported. Additionally, somatic H1 proteins from chicken also cluster by species rather than by type. The lineage of human H1.X subtype is the first to split in the vertebrate group, followed by the differentiation of the testis-specific subtype from mammals (H1t), which is the fastest-evolving histone class, and its synthesis may depend on additional factors to those related with RD and RI expression (Drabent, Kardalinou, and Doenecke 1991), and by the amphibian H1 lineage.

Finally, the avian, fish, and mammalian somatic lineages are differentiated. The divergence of mammalian H1.1 to H1.5 paralogs took place about 390 ± 90 MYA on average, and the time for the divergence of the whole set of genes (H1.1 to H1.5 and H1t) was estimated at about 406 ± 80 MYA (Ponte, Vidal-Taboada, and Suau 1998).

H1 Nucleotide Evolution

An additional phylogeny for H1 genes was reconstructed from 146 nucleotide-coding sequences belonging to 55 species, shown in figure 2. It is important to note that our attention focuses more on the phylogenetic tree reconstructed from amino acid sequences because the topology obtained using nucleotide sequences is not very reliable, given that many gene comparisons within and between species are close or have even reached the saturation level. Although H1 is the least-conserved histone class, most of the observed nucleotide divergence is presented as synonymous variation, both within and between species (fig. 2).

The presence of paralogous RD H1 genes located in close proximity on a chromosome in human and mouse genomes allows us to independently determine whether these genes undergo interlocus recombination or gene conversion. If this is the case, the extensive interlocus exchange would homogenize H1 sequences, resulting in a high sequence similarity among paralogs. To test this hypothesis, we have estimated the average numbers of nucleotide differences per site (p) among H1.1 to H1.5 paralogs in each species and also between orthologs from both species. The extent of p ranges from 0.208 ± 0.014 to 0.332 ± 0.018 substitutions per site (humans) and from 0.136 ± 0.013 to 0.309 ± 0.018 substitutions per site (mouse), with overall mean values of 0.266 ± 0.017 and 0.223 ± 0.015, respectively. These values are greater than those estimated between human and mouse orthologs, which reach a peak value in the case of the H1.1 subtype (0.266 ± 0.017), followed by H1.3 (0.207 ± 0.017), H1.5 (0.202 ± 0.014), H1.2 (0.186 ± 0.015), and H1.4 (0.139 ± 0.013) (fig. 3*A*). Our results show that mouse paralogs, which are clustered on chromosome 13, are more closely related to their human orthologs, which are clustered on the human chromosome 6 (fig. 3*B*). As for the case of p, the average values of p_S range from 0.590 ± 0.040 to 0.712 ± 0.038 substitutions per site between human paralogs and from 0.276 ± 0.033 to 0.597 ± 0.040 substitutions per site between mouse paralogs. These values did not differ significantly from those obtained in the comparisons between orthologs, where the highest level of silent divergence was found in the case of H1.1 (0.687 ± 0.038), followed by H1.5 (0.639 ± 0.038), H1.3 (0.569 ± 0.038), H1.2 (0.533 ± 0.039), and H1.4 (0.435 ± 0.039) (fig. 2). When comparing these values with the nonsynonymous differences, we always found that p_S is significantly greater than p_N ($P < 0.001$, Z-test [table 1]).

Although H1 nucleotide sequences diverge extensively through silent substitutions, H1 genes from the same species do not necessarily cluster together in the phylogenies based on synonymous differences (fig. 2) and total nucleotide differences (fig. 3*B*). In general, the extent of synonymous differences between H1 genes was very high, and the range of p_S values was nearly the same for both within species and between related species (table 1). Additionally, the comparisons between representative RD H1 sequences from different eukaryotic kingdoms reveal that genes from a species are no more closely related to each other than they are to genes from species belonging to very different eukaryotic kingdoms (table 2). For example, it is significant that the average synonymous divergence between human H1.1 and H1.5 paralogs is about 0.691 ± 0.041 substitutions per site, which is roughly the same as the silent divergence observed between human H1.1 and fungi H1 genes ($p_S = 0.676 \pm 0.040$).

However, there was a case where intraspecies sequences were closely related to each other. Chicken H1 genes show relatively low p_S values in intraspecific comparisons and also when compared with duck H1 genes, although in this case, they are significantly greater than the magnitude of p_N ($P < 0.001$, Z-test [table 1]).

If the H1 histone multigene family has evolved according to the birth-and-death model of evolution, pseudogenes may have been generated. By comparing the nucleotide differences between pseudogenes and functional genes with the intraspecific nucleotide variation, it is likely that putative pseudogenes identified for *C. elegans* and *A. thaliana* have emerged quite recently because of their low divergence values and relatively short branches in the phylogeny. However, the previously reported *X. laevis* pseudogenes (Turner et al. 1983) and the putative pseudogene identified for *L. esculentum* seem to be older, given their significant sequence divergence with functional H1 genes and longer branch lengths (fig. 2 and table 3).

Evolution of the Replication-Independent H1 Subtypes

The lineage of RI H1 proteins from vertebrates seem to arise a little later than the RD subtypes (fig. 1), showing a split that gives rise to two lineages early in their evolution. One of them gathers the *orphon* H1 proteins from mussels, which are finally differentiated in the H1s organized in clusters containing only H1 proteins and in the H1s present in the repetitive units (Eirín-López et al. 2002, 2004). The second lineage gives rise to the vertebrate differentiation-specific subtypes, gathering the H1° replacement subtypes and the H5 subtypes. A very interesting feature presented by the RI H1 subtypes comes from their long-term evolutionary pattern. RI H1s again cluster by type instead of by species, suggesting that they are more closely related between than within species (figs. 1–3), showing high numbers of synonymous nucleotide differences per site ($p_S = 0.474 \pm 0.017$ on average), which in all cases are significantly greater than the numbers of nonsynonymous nucleotide differences ($P < 0.001$, Z-test [table 1]).

Discussion

Evolutionary Scenario of H1 Genes

The H1 histone multigene family encodes multiple isoforms, including replication-dependent and replication-independent subtypes. The genes coding for the H1.1 to H1.5 and H1t human subtypes are clustered together with

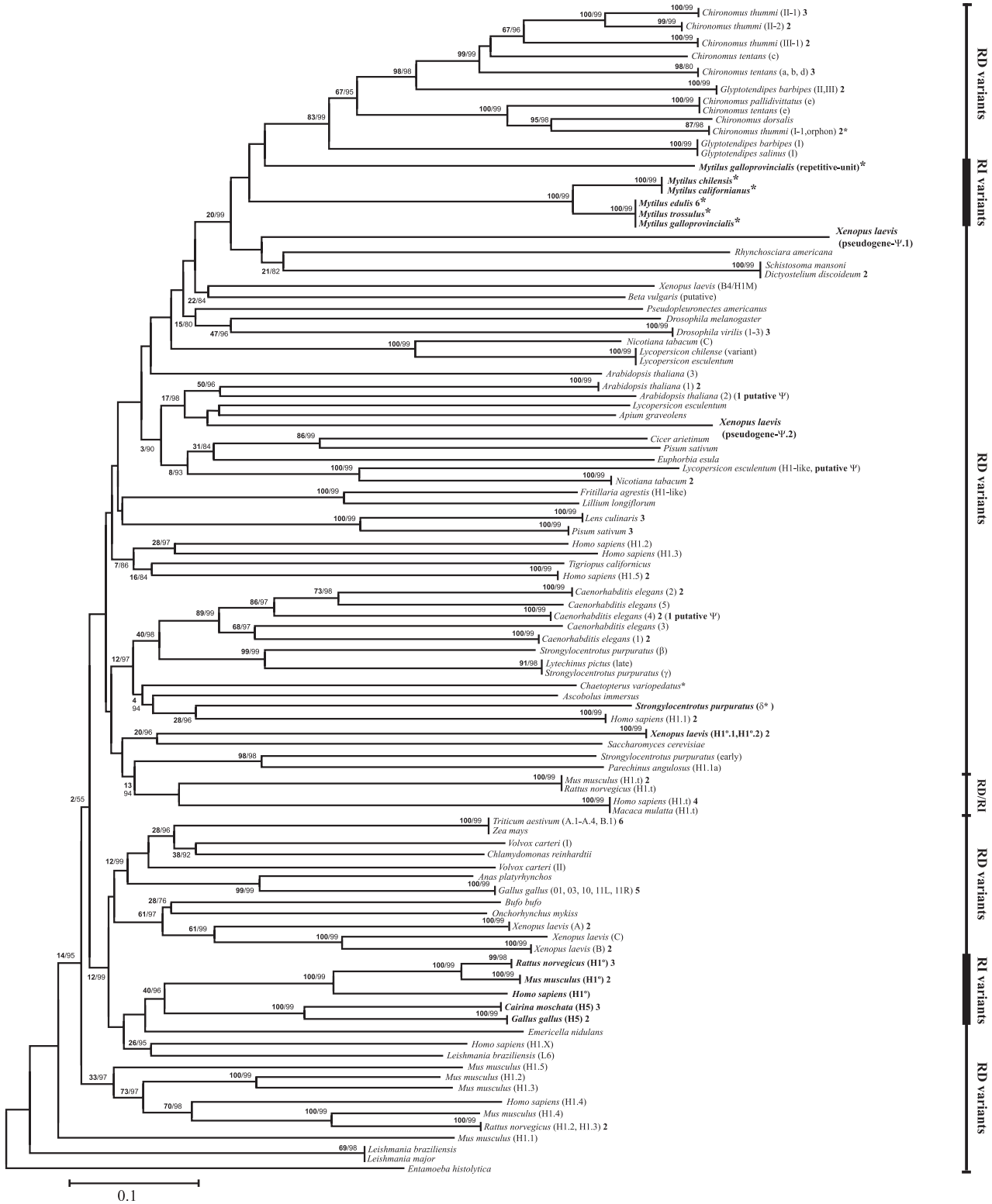

FIG. 2.—Phylogenetic neighbor-joining tree of H1 complete nucleotide-coding sequences using the number of synonymous nucleotide differences per site (p_S) computed by means of the modified Nei-Gojobori method (*p*-distance). BP values (boldface) followed by CP values (normal) are placed in the corresponding nodes and only shown when a value is greater than 50% of the 1,000 replicates. The H1 subtypes and the number of coding sequences are indicated near the corresponding species in parentheses and in boldface, respectively. Pseudogenes are referred to as Ψ in boldface and expression patterns are indicated as in figure 1.

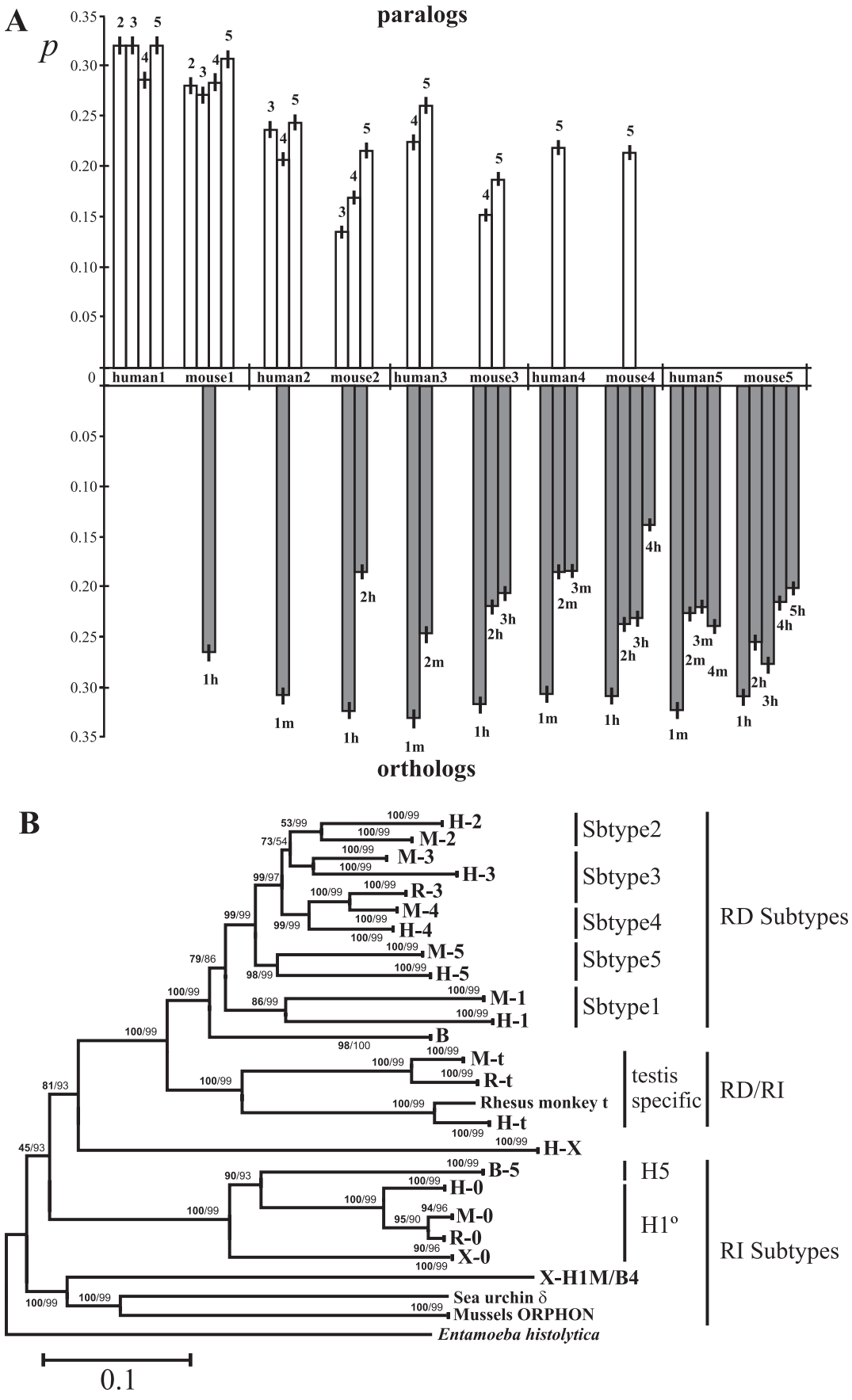
A
paralogs
p
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
human1
mouse1
human2
mouse2
human3
mouse3
human4
mouse4
human5
mouse5
orthologs
B
H-2
M-2
M-3
H-3
R-3
M-4
H-4
M-5
H-5
M-1
H-1
B
M-t
R-t
Rhesus monkey t
H-t
H-X
B-5
H-0
M-0
R-0
X-0
X-H1M/B4
Sea urchin δ
Mussels ORPHON
Entamoeba histolytica
Sbtype2
Sbtype3
Sbtype4
Sbtype5
Sbtype1
RD Subtypes
testis specific
RD/RI
H5
H1°
RI Subtypes
0.1

Table 1
Average Numbers of Synonymous (p_S) and Nonsynonymous (p_N) Nucleotide Differences per site and Average Transition/Transversion Ratio (R) in H1 Genes from Representative Vertebrate, Invertebrate, Plant, and Fungus H1 Genes[a]

	p_S (SE)	p_N (SE)	R[a]		p_S (SE)	p_N (SE)	R[a]
VERTEBRATES				PLANTS			
Chicken	0.155 (0.018)	0.041 (0.006)	1.2**	*Volvox cartei*	0.585 (0.035)	0.319 (0.026)	0.6**
Human (genes 1–5)	0.557 (0.016)	0.120 (0.012)	1.2**	*Arabidopsis* (genes 1–3)	0.545 (0.021)	0.333 (0.018)	0.7**
Mouse (genes 1–5)	0.472 (0.021)	0.129 (0.013)	1.0**	Tobacco	0.707 (0.043)	0.432 (0.027)	0.7**
Xenopus laevis (genes A–C)	0.209 (0.022)	0.087 (0.010)	1.1**	*Lens culinaris*	0.183 (0.010)	0.000 (0.000)	4.0**
Chicken/Duck	0.302 (0.028)	0.047 (0.008)	1.3**	Tomato	0.776 (0.044)	0.408 (0.021)	0.8**
Mouse/Rat (gene 3)	0.335 (0.027)	0.052 (0.009)	1.2**	Wheat	0.213 (0.017)	0.076 (0.007)	0.5**
Mammals	0.566 (0.013)	0.165 (0.011)	1.0**	Pea	0.392 (0.021)	0.189 (0.014)	0.9**
Xenopus/Bufo	0.409 (0.035)	0.164 (0.016)	1.0**	*L. esculentum/L. chilense*	0.476 (0.037)	0.207 (0.026)	0.8**
INVERTEBRATES				*V. carteri/C. reinhardtii*	0.501 (0.036)	0.252 (0.024)	0.5**
Chironomus tentans	0.346 (0.021)	0.095 (0.010)	1.1**	FUNGI			
Drosophila	0.355 (0.019)	0.135 (0.012)	0.8**	Fungi	0.680 (0.025)	0.440 (0.022)	0.6**
Glyptotendipes barbipes	0.453 (0.028)	0.207 (0.017)	0.8**				
Mytilus galloprovincialis	0.627 (0.045)	0.155 (0.023)	1.1**	RD subtypes	0.643 (0.010)	0.355 (0.031)	0.6**
S. purpuratus	0.672 (0.020)	0.402 (0.021)	0.6**	RI subtypes	0.474 (0.017)	0.135 (0.011)	1.2**
C. elegans (genes 1–5)	0.409 (0.021)	0.187 (0.013)	1.0**	H1°	0.427 (0.018)	0.076 (0.009)	1.3**
C. thummi/C. tentans	0.376 (0.034)	0.117 (0.010)	1.1**	H5	0.199 (0.022)	0.045 (0.009)	1.4**
D. melanogaster/D. virilis	0.644 (0.022)	0.238 (0.015)	0.8**	Invertebrate orphons	0.280 (0.014)	0.086 (0.007)	0.8**
S. purpuratus/L. pictus	0.495 (0.036)	0.208 (0.022)	0.6**	H1t tissue-specific	0.352 (0.021)	0.142 (0.013)	1.5**

[a] $p_S > p_N$ in all Z-test comparisons; the significance levels are indicated by *($P < 0.05$) and by **($P < 0.001$). S.E. indicates standard errors calculated by the bootstrap method with 1,000 replicates.

core histones in the chromosomes 6 (major cluster) and 3 (minor cluster) (Albig et al. 1997). In mouse, they are located in chromosomes 13 (major cluster) and 3 (minor cluster) (Wang et al. 1997). The human H1° subtype is present as a single-copy gene in chromosome 22, whereas mouse H1° is located in chromosome 15, which, curiously, is in part syntenic to the human chromosome 22 (Brannan et al. 1992). Under concerted evolution, there would be extensive homogenization among paralogs in close proximity on a chromosome (DeBry and Marzluff 1994). The topologies obtained in the phylogenetic trees (figs. 1–3) show that human and mouse H1 sequences intermingle extensively and are clustered by H1 type, indicating that they are more closely related between than within species and that these genes have not been subject to any significant interlocus homogenization of sequences within either of the two species. In this case, the functional roles of somatic H1 isoforms in chromatin condensation and regulation of gene expression are very important constraints in maintaining the protein structure associated with a concrete and critical function. These results agree with the birth-and-death model, where protein homogeneity is maintained by strong purifying selection, and alleles from different loci are expected to form different clusters (Nei and Hughes 1992; Nei, Gu, and Sitnikova 1997; Nei, Rogozin, and Piontkivska 2000).

If there is an evolution through a rapid process of interlocus recombination or gene conversion, both p_S and p_N would acquire similar values. Our results show that the extent of p_S is always significantly greater than that of p_N in comparisons both within and between species (table 1), suggesting an extensive silent divergence among H1 genes. Additionally, most of the estimated intraspecific p_S values are as high as the p_S values obtained between species, even those belonging to different eukaryotic kingdoms (table 2). These results, rather than an important effect of interlocus recombination, best fit the birth-and-death model, where the nucleotide divergence among members of the multigene family will be observed primarily at the synonymous level and pairs of genes that were duplicated recently are expected to be closely related or even identical (Nei, Rogozin, and Piontkivska 2000). The only exception to this observation was presented by chicken H1 genes, which show high sequence similarity. A possible explanation for this high level of similarity could involve (1) the high GC content in these genes (GC at third codon positions is 84% to 91% in chicken H1 genes), (2) a recent gene duplication within a short period of time (not enough time could have elapsed to allow for the accumulation of nucleotide substitutions), or (3) a gene conversion event, which could not be completely discarded in this case.

As mentioned above, under the birth-and-death model of evolution with strong purifying selection, some of the duplicated genes may become pseudogenes. Until now, the only example of H1 pseudogenes was described in

←

FIG. 3.—(*A*) Average numbers of total nucleotide differences per site among human and mouse H1 paralogs (upper axis) and between human and mouse H1 orthologs (lower axis) using uncorrected *p*-distances. The five H1 somatic subtypes are indicated by their numbers (1 to 5) and referred to human (h) and mouse (m). Bars indicate the standard errors computed by the bootstrap method (1,000 replicates). (*B*) Phylogenetic tree of vertebrate RD and RI H1 complete nucleotide-coding sequences. Uncorrected *p*-distances were used and BP and CP values are indicated as in figures 1 and 2. Species abbreviations are as follows: B, bird; H, human; M, mouse; R, rat; and X, *Xenopus*. Different H1 subtypes are indicated in the right margin of the tree.

Table 2
Numbers of Synonymous Nucleotide Differences per Site (p_S, Below Diagonal) and Standard Errors (S.E., Above Diagonal) in RD H1 Genes of Vertebrates, Invertebrates, Plants, and Fungi

	H-1	H-5	M-3	M-5	X-C	D	Myt	C-1	A-2	E
H-1	—	0.041	0.041	0.042	0.041	0.036	0.040	0.038	0.039	0.040
H-5	0.691	—	0.037	0.038	0.037	0.035	0.042	0.036	0.033	0.041
M-3	0.715	0.617	—	0.037	0.040	0.035	0.040	0.035	0.032	0.038
M-5	0.646	0.669	0.474	—	0.038	0.035	0.041	0.037	0.033	0.039
X-C	0.818	0.736	0.676	0.767	—	0.037	0.045	0.039	0.034	0.045
D	0.752	0.732	0.748	0.758	0.766	—	0.039	0.039	0.035	0.040
Myt	0.804	0.794	0.764	0.865	0.773	0.744	—	0.043	0.040	0.040
C-1	0.726	0.654	0.644	0.644	0.635	0.717	0.687	—	0.036	0.041
A-2	0.688	0.771	0.772	0.715	0.710	0.755	0.779	0.752	—	0.038
E	0.676	0.674	0.664	0.662	0.590	0.770	0.733	0.684	0.754	—

NOTE.—H1-1, human H1.1; H-5, human H1.5; M-3, mouse H1.3; M-5, mouse H1.5; X-C, *Xenopus laevis* H1C; D, *Drosophila melanogaster* H1; Myt, *Mytilus galloprovincialis* H1; C-1; *Caenorhabditis elegans* H1.1; A-2, *Arabidopsis thaliana* H1-2; E, *Emericella nidulans* H1. Standard errors were computed using the bootstrap method (1,000 replicates).

Xenopus laevis (Turner et al. 1983). In our screening of the databases, we did not find any RD or RI truncated H1 sequences. Nevertheless, it was possible to define putative pseudogenes in *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Lycopersicon esculentum*, based on their unusual sequence features. The absence of significant differences from functional H1 genes and the moderate lengths of the branches in the phylogeny (table 3 and fig. 2) suggest a recent loss of function in the case of putative H1 pseudogenes from *C. elegans* and *A. thaliana*, as was shown by Ota and Nei (1994) for immunoglobulin V_H genes. Pseudogenes from *X. laevis* and the putative pseudogene from *L. esculentum*, which show significant differences with functional genes, seem to be otherwise quite old (table 3). In the case of *X. laevis*, pseudogenes show the longest branch lengths in the phylogeny (fig. 2), which agrees with the birth-and-death model and suggests that neither intergenic gene conversion nor unequal crossing-over play major roles in homogenizing these genes (Ota and Nei 1994). Because H1 histones are less conserved compared with core histones, to clearly identify pseudogenes becomes a very problematic issue. Nevertheless, the presence of pseudogenes is not an absolute "must-be" condition of the birth-and-death model of evolution if the remaining assumptions are satisfied (Nei and Hughes 1992; Nei, Gu, and Sitnikova 1997; Nei, Rogozin, and Piontkivska 2000).

Table 3
Pseudogene and Functional H1 Nucleotide Divergences Using Uncorrected *p*-Distances

	Divergence *p*-Distance (S.E.)	
Putative Pseudogene	Pseudogene vs. Functional	Average Functional Genes
Xenopus laevis (Ψ.1)	0.798 (0.022)	0.309 (0.022)*
Xenopus laevis (Ψ.2)	0.723 (0.021)	0.309 (0.022)*
Caenorhabditis elegans	0.481 (0.016)	0.341 (0.013)
Arabidopsis thaliana	0.548 (0.018)	0.481 (0.017)
Lycopersicon esculentum	0.615 (0.016)	0.382 (0.018)*

NOTE.—Asterisk (*) indicates significance level of $P < 0.001$ in Z-test comparisons between pseudogene versus functional genes. Standard errors (S.E.) were computed by the bootstrap method (1,000 replicates) and are indicated in parenthesis.

The presence of clustered H1 RD variants and solitary H1 RI variants allows us to determine whether, as predicted by the concerted evolution model, clustered genes show evidence of interlocus recombination more often than solitary genes (Ohta 1983). Our results show that protein homogeneity is also maintained by strong purifying selection in RI subtypes, which keep their identities and are more closely related between species (figs. 1–3). In this case, the presence of functional constraints would also account for the homologies observed among RI proteins from vertebrates. At the nucleotide level, there is also an extensive silent divergence both within and between species, which is always significantly greater than the nonsilent divergence (table 1). Again, the presence of a significant effect of interlocus recombination at the protein level in RI H1 histones seems unlikely, being probable that RI variants, as RD variants, evolve following the birth-and-death model of evolution with strong purifying selection.

Origin and Long-Term Evolution of RI *orphon* H1 Genes

The phylogenies reconstructed in the present work show that neither the *orphon* H1 variant from the midge *Chironomus thummi* (Hankeln and Schmidt 1993) nor the polyadenylated H1 gene from the annelid *Chaetopterus variopedatus* (del Gaudio et al. 1998) are included in the monophyletic group gathering the RI variants (figs. 1–3). An RI status was proposed for the cases cited above on the basis of their solitary genomic organization, analysis of promoter regions, and presence of putative polyadenylation signals, but except for the sea urchin H1δ histone (Lieber et al. 1988), this latter feature has been inferred from nucleotide sequences rather than by expression analyses. The results of our Northern blotting experiments on mussel *Mytilus galloprovincialis* RNA show the presence of polyadenylated H1 transcripts, which together with previous evidence (Eirín-López et al. 2002, 2004), will definitively demonstrate the RI status for a fraction of H1 genes in mussels.

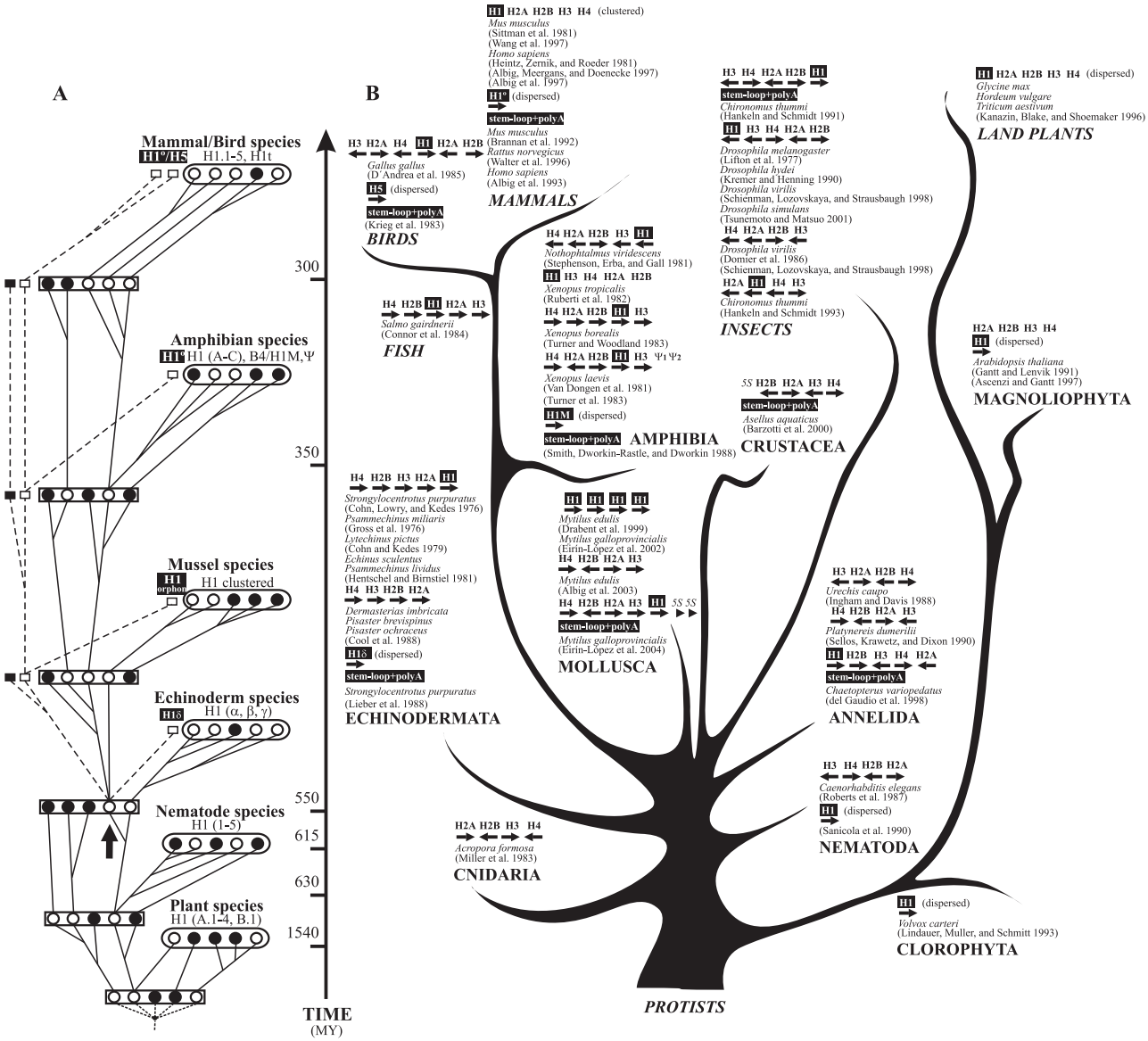

FIG. 4.—(*A*) Scheme of the birth-and-death model of evolution applied to the case of the H1 histone multigene family, adapted from figure 1 in Nei, Gu, and Sitnikova (1997). Open and black circles indicate functional and nonfunctional (pseudogenes) RD H1 genes, respectively, indicating the isoforms differentiated for several taxonomic groups above. The black arrow indicates the event of transposition of an RD H1 gene occurred before vertebrate and invertebrate differentiation, giving rise to solitary RI H1 genes, indicated by open (functional) and black (nonfunctional) boxes. RI variants are also indicated for each group, and they would continue their birth-and-death evolution (dotted lines) in a new physical location and independently from RD variants (solid lines). (*B*) Schematic "tree of life" showing the organization and polarity (indicated by black arrows) of H1 histone genes and core histone genes in model organisms as in many other genomes. This figure parallels figure 4*A* and shows precisely the events of duplication, deletion, and transposition involved in the birth-and-death evolution of H1 genes along the evolutionary scale. Special attention is paid to whether H1 genes are in the major repetitive units or solitary in the genome and whether they show RI features as polyadenylation signals (a key feature in the evolution of RI variants, highlighted with black boxes). The divergence times of the groups were assigned as indicated by Feng, Cho, and Doolittle (1997), and by Peterson et al. (2004) in the case of the origin of bilateria.

An *orphon* origin was hypothesized to explain the evolutionary origin of the RI H1 subtypes from vertebrates, where the exclusion of these genes from the main histone repetitive units and consequently from the interlocus recombination or concerted evolution events, would account for the presence of this differentiation-specific subtypes solitary in the genome (Schulze and Schulze 1995). If the effect of concerted evolution on the long-term evolution of both RD and RI H1 subtypes is not significant, as revealed in the present work, it is then necessary to revisit this *orphon* origin hypothesis to fit it into the birth-and-death model of evolution. A brief scheme of the model of birth-and-death evolution (Nei, Gu, and Sitnikova 1997) is adapted to the concrete case of H1 genes in figure 4*A*. Following this model, the different H1 isoforms may have been generated by recurrent gene duplication/deletion events. Functional H1 proteins would evolve under a strong purifying selection determined by their critical structural and functional roles, which would be already operating at the time of divergence of the RI H1 genes before the

differentiation between vertebrates and invertebrates, about 815 MYA (Feng, Cho, and Doolittle 1997). At the nucleotide level, H1 genes may diverge extensively through synonymous substitution events, being DNA sequences of different gene family members very different both within and between species (Nei and Hughes 1992; Nei, Gu, and Sitnikova 1997). This events proposed theoretically in figure 4*A* are precisely shown by real data in figure 4*B*. This "tree of life" shows the organization of H1 and core histone genes in model organisms as well as in many other genomes, indicating the modifications in histone organization with special attention to whether H1 genes are in the major repetitive units or solitary in the genome and if they show RI features as polyadenylation signals. The next step after the duplication events would involve the transposition of RI H1 genes to a solitary location in the genome, where they would continue their evolution in a new physical location and where new genes and pseudogenes would be generated. The presence of transposition and inversion events is very common in histone evolution, as revealed by the different histone gene orientations in the DNA strands, and a similar pattern of duplication and transposition events has been postulated to explain the long-term evolution of the multigene families of the vertebrate immune system (Sitnikova and Nei 1998).

The final step of the process would involve the acquisition of both an RI gene expression pattern and a concrete function by these *orphon* variants from invertebrates. Although this issue is very well documented in the case of vertebrates, a RI status for several invertebrate H1 genes has been inferred based only on putative sequence features whose functionality was not fully demonstrated. Only expression analysis of these "putative" RI H1 genes from invertebrates will definitively clarify whether they follow an RI expression pattern and if so, whether these polyadenylated transcripts are ubiquitous, circumscribed to certain tissues, or expressed in specific developmental stages. An additional interesting question concerns the analysis of the H1 promoter regions, which were studied in mussel and sea urchin H1 genes together with vertebrate RI H1 genes (H1°/H5), finding significant homologies among them (Eirín-López et al. 2002, 2004). These results are in agreement with those reported in the present work, where RI subtypes (including mussel H1 genes) cluster together by type and not by species. The case of the tissue-specific H1t histone is more complex because its synthesis may depend on different factors than those related with RD and RI expression, but their promoter regions (Drabent, Kardalinou, and Doenecke 1991) and nucleotide coding regions again cluster by type and not by species.

In the present work, we have shown that although the members of the H1 histone multigene family encode a set of highly conserved proteins, they do not evolve in a concerted manner. The diversification of the H1 isoforms is enhanced primarily by mutation and selection, where genes are subject to birth-and-death evolution with strong purifying selection. This model is able to explain not only the diversification of RD H1 genes but also the origin and long-term persistence of *orphon* RI H1 subtypes in the genome. It is likely that H1 genes have experienced a faster birth rate and an apparently slower death rate compared with H3 and H4 families (Piontkivska, Nei, and Rooney 2002; Rooney, Piontkivska, and Nei 2002), given the greater diversification of the H1 isoforms and the few pseudogenes detected. Nevertheless, the long-term evolution of the H1 genes may have paralleled that of core histone genes to maintain a coordinate regulation (Peretti and Khochbin 1997). It seems that multigene families such as histones, which have evolved to produce a large quantity of the same gene product, also evolve at long-term following the birth-and-death model of evolution.

Supplementary Material

Supplementary table shows GenBank accession numbers for H1 nucleotide and protein sequences used in this study.

Acknowledgments

We thank Juan Ausió, Helen Piontkivska, Alejandro Rooney, Nandy Ruiz, and Lucas Sánchez for their valuable comments on an earlier version of this paper. This work was funded by a grant from the PGIDT (10PX110304) given to J.M. and by a predoctoral FPU fellowship from the Spanish government awarded to J.M.E.-L.

Literature Cited

Albig, W., P. Kioschis, A. Poutska, K. Meergans, and D. Doenecke. 1997. Human histone gene organization: nonregular arrangement within a large cluster. Genomics **40**: 314–322.

Albig, W., T. Meergans, and D. Doenecke. 1997. Characterization of the H1.5 genes completes the set of human H1 subtype genes. Gene **184**:141–148.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Brannan, C. I., D. J. Gilbert, J. D. Ceci, Y. Matsuda, V. M. Chapman, J. A. Mercer, H. Eisen, L. A. Johnston, N. G. Copeland, and N. A. Jenkins. 1992. An interspecific linkage map of mouse chromosome 15 positioned with respect to the centromere. Genomics **13**:1075–1081.

Chabouté, 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.

Coen, E., T. Strachan, and G. A. 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.

D'Andrea, R., L. S. Coles, C. Lesnikowski, L. Tabe, and J. R. E. Wells. 1985. Chromosomal organization of chicken histone genes: preferred association and inverted duplications. Mol. Cell. Biol. **5**:3108–3115.

DeBry, R. W., and W. F. Marzluff. 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics **138**:191–202.

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 polychaete worm *Chaetopterus variopedatus*: first record of a H1 histone gene in the phylum annelida. J. Mol. Evol. **46**:64–73.

Dimitrov, S., G. Almouzni, M. Dasso, and A. P. Wolffe. 1993. Chromatin transitions during early *Xenopus* embryogenesis: changes in histone H4 acetylation and in linker histone type. Dev. Biol. **160**:214–227.

Doenecke, D., W. Albig, C. Bode, B. Drabent, K. Franke, K. Gavenis, and O. Witt. 1997. Histones: genetic diversity and tissue-specific gene expression. Histochem. Cell. Biol. **107**: 1–10.

Dover, G. 1982. Molecular drive: a cohesive mode of species evolution. Nature **299**:111–117.

Drabent, B., E. Kardalinou, and D. Doenecke. 1991. Structure and expression of the human gene encoding testicular H1 histone (H1t). Gene **103**:263–268.

Eirín-López, J. M., A. M. González-Tizón, A. Martínez, and J. Méndez. 2002. Molecular and evolutionary analysis of mussel histone genes (*Mytilus* spp.): possible evidence of an 'orphon origin' for H1 histone genes. J. Mol. Evol. **55**:272–283.

Eirín-López, J. M., M. F. Ruiz, A. M. González-Tizón, A. Martínez, L. Sánchez, and J. Méndez. 2004. Molecular evolutionary analysis of the mussel *Mytilus* histone multigene family: first record of a tandemly repeated unit of five histone genes containing an H1 subtype with 'orphon' features. J. Mol. Evol. **58**:131–144.

Felsestein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

Feng, D. F., G. Cho, and R. S. Doolittle. 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028–13033.

Fry, B. G., W. Wüster, R. M. Kini, V. Brusic, A. Khan, D. Venkataraman, and A. P. Rooney. 2003. Molecular evolution and phylogeny of elapid snake venom three-finger toxins. J. Mol. Evol. **57**:110–129.

Gu, X., and M. Nei. 1999. Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. Mol. Biol. Evol. **16**:147–156.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. **41**:95–98.

Hankeln, T., and E. R. Schmidt. 1993. Divergent evolution of an 'orphon' histone gene cluster in *Chironomus*. J. Mol. Biol. **234**:1301–1307.

Hentschel, C. C., and M. L. Birnstiel. 1981. The organization and expression of histone gene families. Cell **25**:301–313.

Isenberg, I. 1979. Histones. Annu. Rev. Genet. **48**:159–191.

Kasinsky, H. E., J. D. Lewis, J. B. Dacks, and J. Ausió. 2001. Origin of H1 histones. FASEB J. **15**:34–42.

Kedes, L. 1979. Histone messengers and histone genes. Annu. Rev. Biochem. **48**:159–191.

Khochbin, S., and A. P. Wolffe. 1994. Developmentally regulated expression of linker-histone variants in vertebrates. Eur. J. Biochem. **225**:501–510.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetic analysis software. Bioinformatics **17**:1244–1245.

Lieber, T., L. M. Angerer, R. C. Angerer, and G. Childs. 1988. A histone H1 protein in sea urchins is encoded by poly(A)$^+$ mRNA. Proc. Natl. Acad. Sci. USA **85**:4123–4127.

Maxson, R., R. Cohn, and L. Kedes. 1983. Expression and organization of histone genes. Annu. Rev. Genet. **17**:239–277.

Nam, J., J. Kim, S. Lee, G. An, H. Ma, and M. Nei. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc. Natl. Acad. Sci. USA **101**:1910–1915.

Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94**:7799–7806.

Nei, M., and A. L. Hughes. 1992. Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. Pp. 27–38 *in* K. Tsuji, M. Aizawa, and T. Sasazuki, eds. Eleventh histocompatibility workshop and conference. Oxford University Press, Oxford, England.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford, England.

Nei, M., I. B. Rogozin, and H. Piontkivska. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc. Natl. Acad. Sci. USA **97**:10866–10871.

Nikolaidis, N., and M. Nei. 2004. Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. Mol. Biol. Evol. **21**:498–505.

Ohsumi, K., and C. Katagiri. 1991. Occurrence of H1-subtypes specific to pronuclei and cleavage stage cell nuclei of anuran amphibians. Dev. Biol. **147**:110–120.

Ohta, T. 1983. On the evolution of multigene families. Theor. Popul. Biol. **23**:216–240.

Ota, T., and M. Nei. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V_H gene family. Mol. Biol. Evol. **11**:469–482.

Peretti, M., and S. Khochbin. 1997. The evolution of the differentiation-specific histone H1 gene basal promoter. J. Mol. Evol. **44**:128–134.

Peterson, K. J., J. B. Lyons, K. S. Nowak, C. M. Takacs, M. J. Wargo, and M. A. McPeek. 2004. Estimating metazoan divergence times with a molecular clock. Proc. Natl. Acad. Sci. USA **101**:6536–6541.

Piontkivska, H., A. P. Rooney, and M. Nei. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol. Biol. Evol. **19**:689–697.

Ponte, I., J. M. Vidal-Taboada, and P. Suau. 1998. Evolution of the vertebrate H1 histone class: evidence for the functional differentiation of the subtypes. Mol. Biol. Evol. **15**: 702–708.

Robertson, H. M. 2000. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. **10**: 192–203.

Rooney, A. P., H. Piontkivska, and M. Nei. 2002. Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. Mol. Biol. Evol. **19**:68–75.

Rzhetsky, A., and M. Nei. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9**: 945–967.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Schienman, J. E., E. R. Lozovskaya, and L. D. Strausbaugh. 1998. *Drosophila virilis* has atypical kinds and arrangements of histone repeats. Chromosoma **107**:529–539.

Schulze, E., and B. Schulze. 1995. The vertebrate linker histones H1°, H5, and H1M are descendants of invertebrate 'orphon' histone H1 genes. J. Mol. Evol. **41**:833–840.

Sitnikova, T. 1996. Bootstrap method of interior-branch test for phylogenetic trees. Mol. Biol. Evol. **13**:605–611.

Sitnikova, T., and M. Nei. 1998. Evolution of immunoglobulin kappa chain variable region genes in vertebrates. Mol. Biol. Evol. **15**:50–60.

Sitnikova, T., A. Rzhetsky, and M. Nei. 1995. Interior-branch and bootstrap tests of phylogenetic trees. Mol. Biol. Evol. **12**:319–333.

Su, C., and M. Nei. 2001. Evolutionary dynamics of T-cell receptor VB gene family as inferred from the human and the mouse genomic sequences. Mol. Biol. Evol. **18**:503–513.

Sullivan, S. A., D. W. Sink, K. L. Trout, I. Makalowska, P. L. Taylor, A. D. Baxevanis, and D. Landsman. 2002. The histone database. Nucleic Acids. Res. **30**:341–342.

Tanaka, M., J. D. Hennebold, J. Macfarlane, and E. Y. Adashi. 2001. A mammalian oocyte-specific linker histone gene H1oo: homology with the genes for the oocyte-specific

cleavage stage histone (CS-H1) of sea urchin and the B4/H1M histone of the frog. Development **128**:655–664.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

Turner, P. C., T. C. Aldridge, H. R. Woodland, and R. W. Old. 1983. Nucleotide sequences of H1 histone genes from *Xenopus laevis*: a recently diverged pair of H1 genes and an unusual H1 pseudogene. Nucleic Acids Res. **11**:4093–4106.

Wang, Z. F., A. M. Sirotkin, G. M. Buchold, A. I. Skoultchi, and W. F. Marzluff. 1997. The mouse histone H1 genes: gene organization and differential regulation. J. Mol. Biol. **271**: 124–138.

Zhang, J., K. D. Dyer, and H. F. Rosenberg. 2000. Evolution of the rodent eosinophil-associated Rnase gene family by rapid gene sorting and positive selection. Proc. Natl. Acad. Sci. USA **97**:4701–4706.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Takashi Gojobori, Associate Editor

Accepted July 9, 2004