Ancient Genome Duplications Did Not Structure the Human *Hox*-Bearing Chromosomes

Austin L. Hughes,[1,3] Jack da Silva,[2] and Robert Friedman[1]

[1]*Department of Biological Sciences, University of South Carolina, Columbia, South Carolina 29208, USA;* [2]*Department of Biology, East Carolina University, Greenville, North Carolina 27858, USA*

The fact that there are four homeobox (*Hox*) clusters in most vertebrates but only one in invertebrates is often cited as evidence for the hypothesis that two rounds of genome duplication by polyploidization occurred early in vertebrate history. In addition, it has been observed in humans and other mammals that numerous gene families include paralogs on two or more of the four *Hox*-bearing chromosomes (the chromosomes bearing the *Hox* clusters; i.e., human chromosomes 2, 7, 12, and 17), and the existence of these paralogs has been taken as evidence that these genes were duplicated along with the *Hox* clusters by polyploidization. We tested this hypothesis by phylogenetic analysis of 42 gene families including members on two or more of the human *Hox*-bearing chromosomes. In 32 of these families there was evidence against the hypothesis that gene duplication occurred simultaneously with duplication of the *Hox* clusters. Phylogenies of 14 families supported the occurrence of one or more gene duplications before the origin of vertebrates, and of 15 gene duplication times estimated for gene families evolving in a clock-like manner, only six were dated to the same time period early in vertebrate history during which the *Hox* clusters duplicated. Furthermore, of gene families duplicated around the same time as the *Hox* clusters, the majority showed topologies inconsistent with their having duplicated simultaneously with the *Hox* clusters. The results thus indicate that ancient events of genome duplication, if they occurred at all, did not play an important role in structuring the mammalian *Hox*-bearing chromosomes.

Ohno (1970) was the first to suggest that one or more rounds of genome duplication by polyploidization played an important role in the early evolution of vertebrates. Recently, this view has achieved wide popularity among vertebrate developmental biologists and immunologists (Lundin 1993; Sidow 1996; Kasahara et al. 1997). According to a widely cited version of this hypothesis, there were two rounds of polyploidization, one occurring before the divergence of Agnatha (jawless vertebrates) and the other just after (the 2R hypothesis; Sidow 1996). Despite the popularity of this hypothesis, Skrabanek and Wolfe (1998), reviewing data available at that time, concluded that no substantial evidence in support of the 2R hypothesis was yet available.

Sidow (1996) adduced in support of the 2R hypothesis the fact that a number of gene families are known to have four members in vertebrates and one or two in *Drosophila*. However, Hughes (1999a) noted that such a pattern supports the 2R hypothesis only if two conditions are met: (1) The vertebrate members of the family can be shown to have duplicated within the vertebrate lineage; and (2) the phylogeny of the gene family shows a specific topology; namely, that of two clusters of two genes, a topology described as (AB) (CD). To test these predictions, Hughes (1999a) examined the gene families of developmentally important proteins having four members in vertebrates, the very families cited in support of the 2R hypothesis by Sidow (1996). In five families, the phylogeny supported duplication of the vertebrate genes before the divergence of deuterostomes and protostomes, and in four of these there was statistically significant support for this conclusion (Hughes 1999a). In only one of the remaining eight families was the topology of the form predicted by the 2R hypothesis, and statistical support in this case was not significant. In contrast, in six cases there was significant support for the alternative topology (A) (BCD) (Hughes 1999a). Therefore, the first rigorous test of key predictions of the 2R hypothesis provided no supported for the hypothesis.

In addition to data on gene number, another type of data frequently cited in support of the 2R hypothesis takes the form of lists of paralogous genes mapped to various human chromosomes. For example, Kasahara et al. (1997) provided lists of gene families having paralogous members on two or more of human chromosomes 1, 6, 9, and 19. Similarly, Lundin (1993) presented extensive lists of "possible paralogies" on human chromosomes 2, 7, 12, and 17, the chromosomes that bear the *Hox* clusters. However, presentation of such lists as evidence for genome duplication without phylogenetic analysis of the relevant gene families is

[3]Corresponding author.
E-MAIL austin@biol.sc.edu; FAX (803) 777-4002.
Article published on-line before print: *Genome Res.,* 10.1101/gr.160001. Article and publication are at www.genome.org/cgi/doi/10.1101/gr.160001.

problematic. For example, the lists of Kasahara et al. (1997) include five gene families for which phylogenetic analyses reveal that the relevant paralogous genes were duplicated before the origin of vertebrates. In one of these (ABC transporters), the gene duplication occurred before the divergence of eukaryotes and eubacteria; in two others (proteasome components and hsp70), the gene duplication occurred before the animal fungus divergence; and in two others (NOTCH and cytochrome p450), the duplication occurred before the divergence of deuterostomes and protostomes (Hughes 1998a; Yeager and Hughes 1999). Clearly these genes could not have been duplicated as part of a hypothesized polyploidization event early in vertebrate history.

Moreover, lists of putative paralogs cannot be evidence of past genome duplication unless the genes involved are in fact homologous. For example, Lundin (1993) included as evidence of genome duplication the presence of genes for malate dehydrogenase on human chromosomes 2 and 7. However, these two genes are unrelated, although their products have a similar enzymatic function. Similarly, Lundin (1993) lists as a group of paralogous genes the genes encoding the cytokines interferon β3 (chromosome 2), interleukin-6 (chromosome 7), and interferon γ (chromosome 12). In fact, none of these three is homologous to any other. Nonetheless, there are numerous gene families including paralogs on at least two of the human *Hox*-bearing chromosomes.

In the present study, we used phylogenetic analyses of all available gene families known to include paralogs on at least two of the human *Hox*-bearing chromosomes ($n = 42$) as a test of the hypothesis that the human *Hox*-bearing chromosomes are structured by the two rounds of ancient genomic duplication postulated by the 2R hypothesis. Previous phylogenetic analysis of *Hox* genes supported the hypothesis that these genes duplicated early in vertebrate history, sometime after the chordate lineage diverged from the cephalochordate lineage (Amphioxus) (Zhang and Nei 1996). We tested the prediction that paralogous genes on the human *Hox*-bearing chromosomes duplicated simultaneously with one another and with the *Hox* clusters.

To test this prediction we used four methods: (1) We constructed phylogenetic trees of gene families and used the order of branching within these trees to determine the timing of gene duplication relative to the divergence of major taxonomic groups. Because the method of phylogenetic analysis we used is robust to differences in the rate of evolution in different branches of the tree, this method does not presuppose a constant rate of molecular evolution (molecular clock) (Saitou and Nei 1987; Hughes 1998a). (2) When phylogenetic analyses did not determine the timing of duplication events, we tested the relevant portions of gene families for constancy of the rate of evolution; when the molecular clock hypothesis could not be rejected, we estimated divergence times of paralogous gene pairs. (3) For gene families with members on three or more of the *Hox*-bearing chromosomes, we compared phylogenies to test the prediction that, if these genes were duplicated simultaneously, their phylogenies should be congruent. (4) Given our phylogenies we estimated the minimum number of genetic events (gene duplications, deletions, and translocations) required to explain the distribution of members of each gene family on the human *Hox*-bearing chromosomes under the two alternative hypotheses of whole-genome duplication and independent duplication and translocation of each gene family. Note that only the second of these methods is dependent either on a molecular clock or on the accuracy of calibrations from the fossil record.

RESULTS

Phylogenetic Analyses

Phylogenetic analyses were conducted for 42 gene families having members on one or two of the human *Hox*-bearing chromosomes (Table 1). In 14 families, the phylogenetic trees supported a divergence time for one or more clades of sequences including paralogs on the human *Hox*-bearing chromosomes before the origin of vertebrates (Fig. 1). The *CDK* family (Fig. 2) is an example of such a family. The phylogeny was rooted in the midpoint of the longest internal branch, but the conclusions regarding the relative timing of gene duplications are not dependent on rooting. Human *CDK7* (on chromosome 2) clustered with yeast *KIN28*, and the branch separating this cluster from other human and yeast genes received highly significant (99%) bootstrap support. This implies that *CDK7* duplicated before the divergence of animals and fungi. Similarly, human *CDK4* (on chromosome 12) and *CDK5* (on chromosome 7) cluster with fungal genes apart from other human and fungal genes with high bootstrap support (98% and 96%, respectively). Again, this topology implies that *CDK4* and *CDK5* duplicated before the divergence of animals and fungi. On the other hand, the phylogeny does not rule out duplication of *CDK2* (on chromosome 12) and *CDK3* (on chromosome 17) early in vertebrate history.

For each of the duplications that, according to our phylogenies, occurred before the origin of vertebrates, we give the bootstrap value for the critical branch supporting such a duplication time (Fig. 1). Of 25 such branches, 19 received bootstrap support ≥95% (Fig. 2). Phylogenetic analyses also supported the hypothesis that one pair of paralogs, human *RAD52* (on chromosome 12) and ψ*-RAD52* (on chromosome 2), diverged

Table 1. Gene Families Used in Phylogenetic Analyses

Gene family	Abbreviation	Human proteins (Accession no.) [Chromosomal location of gene]
Acetylcholine receptor	*ACHR*	ACHRD (Q07001) [2], ACHRG (P07510) [2], ACHRB (P11230) [17], ACHRE (Q04844) [17]
Acetyl-coA carboxylase		COA2 (O00763) [12], COA1 (Q13085) [17]
Actin	*ACT*	ACTH (P12178) [2], ACTB (P02570) [7], ACTG (P02571) [17]
Acyl-coA dehydrogenase	*ACAD*	ACAD-L (P28330) [2], ACAD-S (P16219) [12], ACAD-VL (P49748) [17], COA-OXP (Q15067) [17]
ADP-ribosylation factor	*ARF*	ARF5 (P26437) [7], ARF3 (P16587) [12], ARL6 (P49703) [17], ARL7 (P56559) [17]
Anion exchanger	*AE*	AE3 (P48751) [2], AE2 (P04920) [7], AE1 (P02730) [17]
Aquoporin	*AQP*	AQP1 (P29972) [7], AQP2 (P41181) [12]
Arrestin	*ARR*	S-ARR (P10523) [2], ARR2 (P32121) [17]
Brain amiloride-sensitive sodium channel	*BNAC*	BNA2 (U78181) [12], BNA1 (Q16515) [17]
Cyclin-dependent kinase	*CDK*	CDK7 (P50613) [2], CDK5 (Q00535) [7], CDK4 (P11802) [12], CDK3 (Q00526) [17]
Enolase	*ENOL*	ENOLG (P09104) [12], ENOLB (P13929) [17]
ERBB receptor protein-tyrosine kinase	*ERBB*	ERBB4 (Q15303) [2], EGFR (P00533) [7], MET (P08581) [7], ERBB3 (P21860) [12], ERBB2 (P04626) [17]
Even-skipped		EVX2 (Q03828) [2], EVX1 (P49640) [7]
Frizzled		FR1 (AB017363) [2], FR7 (AB017365) [7]
GLI zinc-finger protein	*GLI*	GLI2 (P10070) [2], GLI3 (P10071) [7], GLI1 (P08151) [12]
Glucagon	*GCG*	GCG (P01275) [2], GIP (P09681) [17]
Glucose transporter	*GLUT*	GLUT3 (P11169) [12], GLUT4 (P14672) [17]
G protein-coupled receptor	*GPR*	CCR4 (P30991) [2], IL8RA (P25024) [2], IL8RB (P25025) [2], NK-1R (P25103) [2], GPR37 (Y12476 {7], CKR7 (P32248) [17], SSR2 (P30874) [17]
Guanine nucleotide-binding protein	*GNB*	GNB2 (P11016) [7], GNB3 (P16520) [12]
Hedgehog	*HH*	IHH (Q14623) [2], SHH (Q15465) [7]
Hepatocyte nuclear factor		HNFA (P20823) [12], HNFB (P35680) [17]
Immunoglobulin-related	*IG*	CD28 (P10747) [2], CTL4 (P16410) [2], CD4 (P01730) [12], CD7 (P09564) [17]
Inhibin	*INHB*	INHA (P05111) [2], INHBB (P09529) [2], INHBA (P08476) [7], INHBC (P55103) [12]
Insulin-like growth factor-binding protein	*IGBP*	IGBP2 (P18065) [2], IGBP1 (P08833) [7], IGBP3 (P17936) [7], IGBP4 (P22692) [17]
Integrin α	*INTA*	INTA4 (P13612) [2], INTA6 (P23229) [2], INTAV (P06756) [2], INTA5 (P08648) [12], INTA7 (AF032108) [12], INTAIIB (P08514) [17]
Integrin β	*INTB*	INTB6 (P18564) [2], INTB8 (P26012) [7], INTB7 (P26010) [12], INTB3 (P05106) [17]
Intermediate filament	*IF*	Desmin (P17661) [2], KRT1-18 (P05783) [12], KRT2-1 (P04264) [12], KRT2-2E (P35908) [12], KRT2-3 (PP12035) [12], KRT2-4 (P19013) [12], KRT2-5 (P13647) [12], KRT2-6A (P02538) [12], KRT2-7 (P08729) [12], KRT2-8 (P05787) [12], Peripherin (P41219) [12], KRT1-9 (P35527) [17], KRT1-10 (P13645) [17], KRT1-12 (Q99456) [17], KRT1-13 (P13646) {17], KRT1-14 (P02533) [17], KRT1-15 (P19012) [17], KRT1-16 (P30654) [17], KRT1-17 (Q04695) [17], KRT1-17-2 (P08779) [17], KRT1-19 (P08727) [17], KRT1-HA1 (Q15323) [17], KRT1-HA2 (Q14532) [17], KRT1-HA3 (Q14525) {17], KRT1-HA5 (Q92764) [17]
Myosin light chain		MYL1 (P05976) [2], MYL3 (P06741) [2], MYLE (P12829) [17]
NAB transcriptional regulator		NAB1 (NM_005966) [2], NAB2 (U48361) [12]
NRAMP		NRAMP1 (D50402) [2], NRAMP2 (L37347) [12]
Nuclear hormone receptor	*NHR*	NOT2 (P43354) [2], RARG1 (P13631) [12], ESTR ;(P03372) [12], NOFIP (P22736) [12], VDR (P11473) [12], RARA1 (P10276) [17], THRA1 (P21205) [17], THRA2 (P10827 [17]

Table 1. (*Continued*)

Gene family	Abbreviation	Human proteins (Accession no.) [Chromosomal location of gene]
Neurokinin	*NKN*	TAC1 (NM_012666) [7], TAC3 (NM_013251) [12]
Nitric oxide synthetase	*NOS*	NOS3 (P29474) [7], NOS1 (P29475) [12], NOS2 (P35228) [17]
Olfactory receptor	*OR*	OR7-10.3 (AC004853) [7], OR7-138 (U86278) [7], OR7-140 (U86280) [7], OR7-141 (U86281) [7], OR17-2 (P47882) [17], OR17-4 (U53583) [17], OR17-15 (U86244) [17], OR17-16 (U86245) [17], OR17-24 (P47883) [17], OR17-30 (U86240) [17], OR17-32 (P47885) [17], OR17-40 (U04683) [17], OR17-82 (P47886) [17], OR17-93 (U76377) [17], OR17-130 (U86240) [17], OR17-135 (U86241) [17], OR17-201 (U76377) [17], OR17-207 (P47889) [17], OR17-209 (U53583) [17], OR17-210 (U53583) [17], OR17-219 (P47892) [17], OR17-228 (P47893) [17]
Pancreatic polypeptide/ neuropeptide Y		Neuropeptide Y (P01303) [7], pancreatic polypeptide (P01298) [17]
Peroxidase		TPO (P07202) [2], MPO (P05164) [17]
Proteasome β subunit	*PSMB*	PSMBθ (P49720) [2], PSMBD (P28072) [17]
RAD52	*RAD52*	Ψ-RAD52 (U22171-U22172) [2], RAD52 (U12134) [12]
Ras-related	*RASR*	RAB1A (P11476) [2], RAB6 (P20340) [2], RALB (P11234) [2], RALA (P11233) [7], RAP1B (P09526) [12], RAB13 (P51153) [12], K-RAS2A (P01116) [12], K-RAS2B (P01118) [12]
Sodium channel	*SCN*	SCNA2 (Q99250) [2], SCNA6 (Q01118) [2], SCNA4 (P35499) [17]
Synapyobrevin	*SYB*	SYB1 (P23763) [12], SYB2 (P19065) [17]
Wnt-related	*WNT*	WNT2 (P09544) [7], WNT1 (P04628) [12], WNT10B (O00744) [12], WNT3 (A47536) [17]

after the mammalian radiation; the two human genes clustered together with 100% bootstrap support, apart from mouse and chicken genes (Fig. 2). Based on the number of synonymous nucleotide substitutions per site (Nei and Gojobori 1986) and using 110 million years ago (Mya) as the approximate time of the rodent–primate divergence (Kumar and Hedges 1998), we estimated the duplication of human *RAD52* and ψ-*RAD52* at 33 ± 5 Mya.

Duplication Time Estimates

In the case of phylogenies and portions of phylogenies for which duplication around the time of the *Hox* duplications could not be ruled out, we used the two-cluster test and the method of linearized trees to test the molecular clock hypothesis. Significant lack of rate constancy or incompatibility with previous estimates of divergence times of major vertebrate taxa led to rejection of the molecular clock hypothesis. There were 15 pairs of paralogous genes for which the molecular clock was not rejected; for these, gene duplication times were estimated from linearized trees, using calibrations based on divergence times of vertebrate classes (Fig. 3).

The vertebrate *Hox* clusters are hypothesized to have duplicated at some point between the divergence of vertebrates from nonvertebrate chordates and the divergence of cartilaginous fishes. The latter divergence has been dated at 528 ± 56 Mya, and the former was probably no earlier than 750 Mya. (The latter is a conservative estimate that we used in the absence of a consensus date.) Of the 15 duplication time estimates, only six (*BNAC*, *GLI*, *GLUT*, *HH*, *IG*, and *SCN*) fell within that time window, whereas two others (*NHR* and *SYB*) had standard errors that overlap the window (Fig. 3). Three pairs (*ACT*, *ENOL*, and *NKN*) were estimated to have duplicated well after the time window, and four (*AQP*, *ARR*, *GCG* and *CDK*) well before it (Fig. 3). Note that one of the duplications placed well before the *Hox* duplications is that of *CDK2* and *CDK3*, for which the phylogenetic tree (Fig. 2) could not rule out a duplication early in vertebrate history.

Tests of Phylogenetic Consistency

Even among gene duplications likely to have occurred around the same time as the *Hox* duplications, phylogenetic analyses revealed inconsistencies among their phylogenies and between their phylogenies and the *Hox* phylogeny. These inconsistencies show that not all of these genes could have duplicated simultaneously with each other and with the *Hox* clusters. Figure 4a shows in schematic form the rooted tree of mammalian *Hox* clusters constructed by Zhang and Nei (1996). In this tree, *HOXC* (chromosome 2) and *HOXD* (chromosome 12) cluster together, with significant (95%) bootstrap support, but the branching order of the other *Hox* clusters is not resolved, presumably because the sequences used are short and extraordinarily highly conserved (Fig. 4a; Zhang and Nei 1996; Bailey et al. 1997). The tree of collagen genes closely

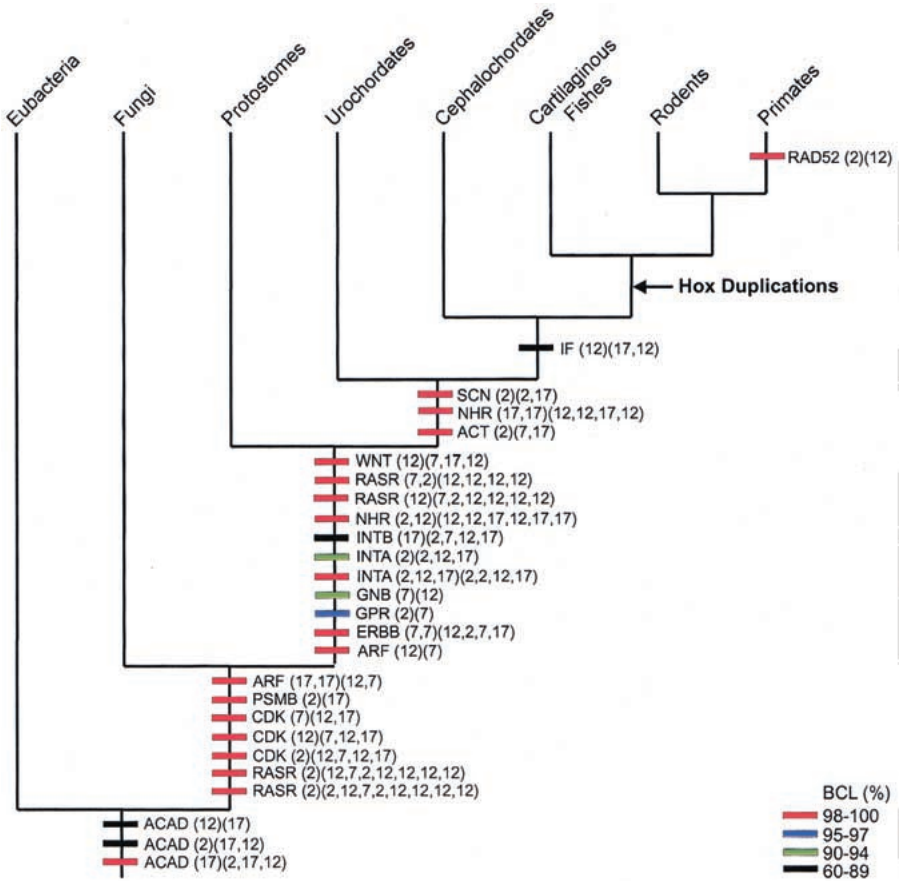

Figure 1 Summary of phylogenetic analyses of families including members on at least two of human *Hox*-bearing chromosomes (2, 7, 12, and 17), which indicated that one or more gene duplications occurred before the origin of vertebrates or after the mammalian radiation. For each family, the timing of duplications is indicated relative to a phylogeny of organisms, and the chromosomal locations of the corresponding human genes are indicated; e.g., the notation "CDK (12) (7,12,17)" indicates that a member of the *CDK* family on human chromosome 12 duplicated from the ancestor of other family members on human chromosomes 7, 12, and 17 before the divergence of fungi from animals. The bootstrap confidence level (BCL) for the internal branches of the phylogenetic trees are color-coded. The time window within which the vertebrate *Hox* clusters are believed to have duplicated is also indicated. Names of gene families are abbreviated as in Table 1.

linked to the *Hox* clusters constructed by Bailey et al. (1997) showed a different topology (Fig. 4b). The collagen genes on chromosomes 7, 12, and 17 formed an unresolved trichotomy, whereas those on chromosome 2 formed an outgroup to them (with 93% bootstrap support; Fig. 4b).

We constructed a phylogenetic tree of members of the *ERBB* family, also closely linked with the *Hox* clusters, and rooted with more distantly related members of the *ERBB* family. The results, summarized in Figure 4c, revealed yet a third topology. In this case, genes on chromosomes 7 and 17 clustered together, with significant bootstrap support; the gene on chromosome 2 branched next; and the gene on chromosome 12 formed an outgroup (again with significant bootstrap support). The hypothesis that *ERBB* paralogs duplicated simultaneously with the linked *Hox* clusters or collagen genes is thus decisively rejected.

Furthermore, certain gene families and subfamilies include members on three of the four human *Hox*-bearing chromosomes. When we constructed phylogenies of these genes, rooted with more distant family members, we could identify two genes as sister groups, with the third as an outgroup to these two (Fig. 5). Of the eight phylogenies, all except that for *INTB* could be accounted for by simultaneous duplication of the genes involved if we assume a phylogeny like that of Figure 4d. In this phylogeny, chromosome 2 and 7 genes cluster together, chromosome 12 genes branch next, and chromosome 17 genes form an outgroup to the others (Fig. 4d). However, this phylogeny is inconsistent with that of the *Hox* clusters, that of the collagen family, or that of *ERBB* (Fig. 4a-c). Thus, even if seven of the eight families and subfamilies shown in Figure 5 did duplicate simultaneously, they did not duplicate along with *Hox*, collagen, or *ERBB* genes.

Number of Genetic Events

Given our phylogenies, we used the maximum parsimony principle to reconstruct the minimal number of character changes (genetic events) required to explain the observed pattern of genes on the human *Hox*-bearing chromosomes under the following alternative hypotheses: (1) the hypothesis of tandem duplication (i.e., that all genes in these families on human chromosomes 2, 7, 12, and 17 arose by independent tandem duplication and translocation events); and (2) the hypothesis of genome duplication (i.e., that two rounds of genome duplication early in vertebrate history contributed to duplication of these genes).

In conducting this analysis, we made conservative assumptions favorable to the genome duplication hypothesis. First, differences among chromosomes with regard to gene order were not considered. If these dif-

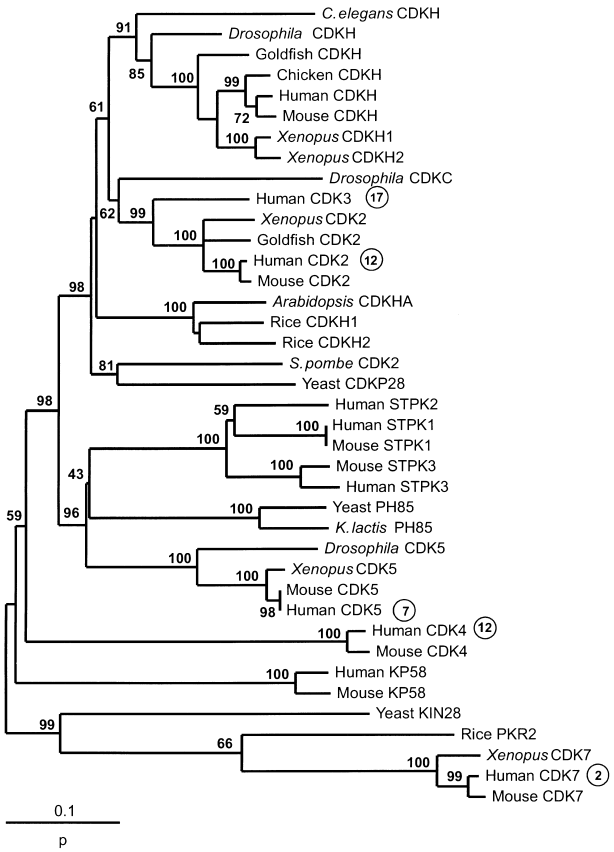

Figure 2 Phylogenetic tree of the cyclin-dependent kinase (*CDK*) family, constructed by the neighbor-joining method, on the basis of the proportion of amino acid difference (p). Numbers on the branches are the bootstrap confidence levels; only values ≥50% are shown. For genes on the human *Hox*-bearing chromosomes, chromosomal locations are indicated by circled numbers. Species included in the phylogeny are the following: Vertebrata: human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), clawed frog (*Xenopus laevis*), goldfish (*Carassius auratus*); Insecta: *Drosophila melanogaster*; Nematoda: *Caenorhabditis elegans*; Plantae: *Arabidopsis thaliana*, rice (*Oryza sativa*); Fungi: yeast (*Saccharomyces cerevisiae*), *Schizosaccharomyces pombe*, *Kluyveromyces lactis*.

ferences were considered, many additional genetic events, involving rearrangement of genes within chromosomes, would have to be postulated under the genome duplication hypothesis but not under the tandem duplication hypothesis. Second, we assumed that the relationship among genes duplicated by polyploidization was as in Figure 4d, which conforms with the phylogeny of a majority of the genes found on three of the four *Hox*-bearing chromosomes (Fig. 5). Third, to explain current gene numbers under the genome duplication hypothesis, it is often necessary to hypothesize loss of paralogs from the *Hox*-bearing chromosomes through either deletion of these genes or their translocation to chromosomes other than the *Hox*-bearing chromosomes (deletion/translocation events). Conservatively, we assumed no more than one such deletion/translocation event per gene family per chromosome.

There were 20 gene families for which the number of reconstructed genetic events differed between the two hypotheses (Table 2). In 14 of these families, the tandem duplication provided a more parsimonious account, and overall the tandem duplication hypothesis required 22 fewer genetic events to explain the observed number and distribution of genes than did the genome duplication hypothesis (Table 2). Thus, given the phylogenies of gene families having members on two or more of the human *Hox*-bearing chromosomes, a hypothesis of multiple independent tandem duplications provides a more parsimonious explanation than does the hypothesis of genome duplication.

DISCUSSION

We used four different approaches to test the predic-

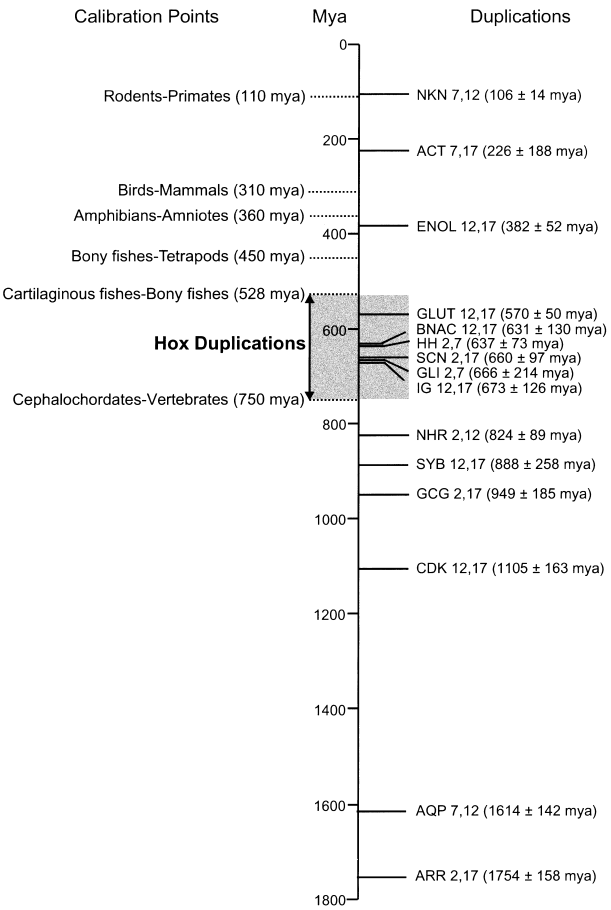

Figure 3 Estimated duplication times (± S.E.) of pairs of paralogous genes on human *Hox*-bearing chromosomes, plotted on a timeline illustrating divergence times of major vertebrate taxa (Kumar and Hedges 1998). Only families for which the hypothesis of a molecular clock could not be rejected were included. The time window within which the vertebrate *Hox* clusters are believed to have duplicated is indicated by shading. Names of gene families are abbreviated as in Table 1.

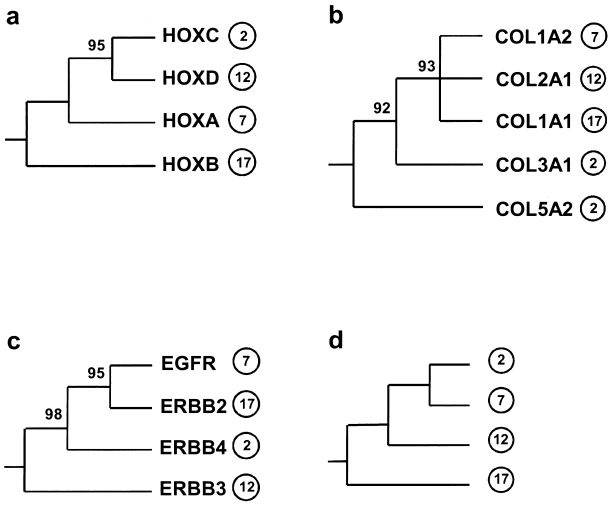

Figure 4 Schematic phylogenies of families having members on all four *Hox*-bearing chromosomes that duplicated early in vertebrate history: (*a*) *Hox* clusters (Zhang and Nei 1996); (*b*) collagen genes (Bailey et al. 1997); (*c*) ERBB family; (*d*) phylogeny consistent with the phylogenies of most three-member families (Fig. 5). Numbers on the branches are the bootstrap confidence levels; only values ≥50% are shown. Chromosomal locations are indicated by circled numbers.

tion that paralogous genes on the human *Hox*-bearing chromosomes duplicated simultaneously early in vertebrate history along with the *Hox* clusters themselves. Combining the results of these methods, a total of 35 families provided evidence regarding the hypothesis of simultaneous duplication, and in 29 of these families the results were inconsistent with that hypothesis. Note that in only four of these families (*AQP*, *ARR*, *ENOL*, and *NKN*) is this evidence dependent on the assumption of a molecular clock. Positive Darwinian selection after gene duplication is one factor that may cause violation of the molecular clock assumption, but it is so far uncertain how widespread this phenomenon is (Hughes 1994, 1999b). In any event, our analysis was conservative in that we dated gene duplications by this method only in cases for which the molecular clock hypothesis could not be rejected using very strict criteria. Even if these four cases are discounted, in a majority of the families analyzed there was strong evidence against the hypothesis of simultaneous duplication early in vertebrate history. Thus, our results falsify a major prediction of the polyploidization hypothesis; namely, that linked paralogous genes arose as a result of simultaneous duplication during polyploidization events (Lundin 1993). Rather, in the case of the human *Hox*-bearing chromosomes, these genes have arisen largely as a result of independent gene duplication and translocation events, scattered at different times over the history of life.

There is evidence that gene numbers of vertebrates are greater than those of invertebrates, including invertebrate chordates. For example, the urochordate *Ciona intestinalis* was recently estimated to have ~15,000 genes (Simmen et al. 1998), as opposed to perhaps ~70,000 in mammals (Miklos and Rubin 1996). Intuitively it might seem that whole-genome duplication by polyploidization would provide a much more parsimonious explanation of such a marked increase in gene number than would an alternative hypothesis involving multiple independent events of tandem duplication and translocation. However, before deciding between these hypotheses in the case of the human *Hox*-bearing chromosomes, it is necessary to determine which hypothesis provides a more parsimonious account of the history of paralogous genes on these chromosomes, given the phylogenies we obtained for the gene families (Fig. 1).

We compared the number of evolutionary events required to explain the observed results under these two hypotheses, under conditions highly favorable to the hypothesis of genome duplication. The results in-

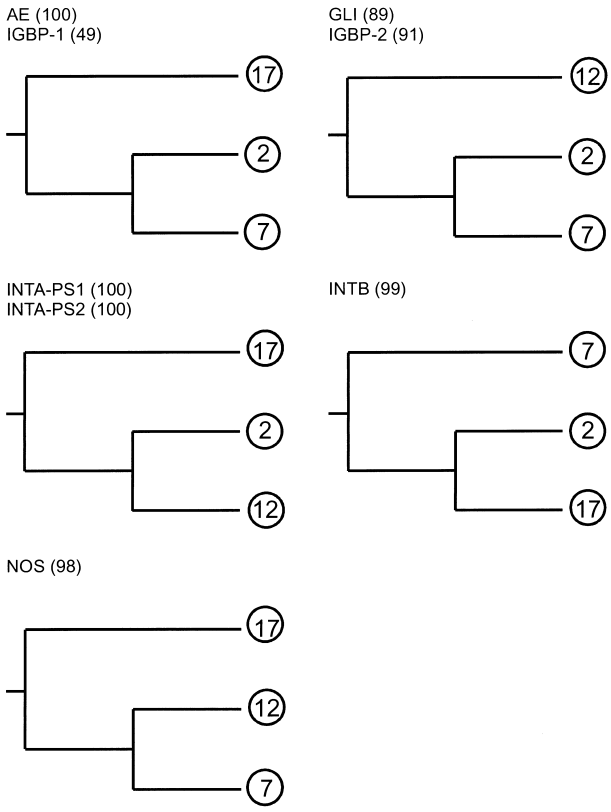

Figure 5 Summary of rooted phylogenies of families of genes including members on three of the four human *Hox*-bearing chromosomes. In each case, the bootstrap confidence level (BCL) of the internal branch is indicated in parentheses. Names of gene families are abbreviated as in Table 1. (INTA-PS1) Vertebrate integrin α chains in the subfamily related to *Drosophila* integrin α-PS1; (INTA-PS2) vertebrate integrin α chains in the subfamily related to *Drosophila* integrin α-PS2 (Hughes 2001). For NOS phylogeny, see Hughes (1998b).

Table 2. Minimum Numbers of Genetic Events (Character Changes) Required to Explain the Occurrence of Gene Family Members on Human Chromosomes 2, 7, 12, and 17

Gene family	Tandem			Genome			
	Dup	Tra	Total	Dup	Tra	Del/Tra	Total
ACT	2	2	4	1	1	4	6
ACAD	3	2	5	3	2	4	9
ARF	3	2	5	3	2	4	9
AE	2	2	4	0	0	1	1
CDK	4	3	7	3	3	4	10
ERBB	5	3	8	3	3	4	9
GLI	2	2	4	0	0	1	1
GNB	1	1	2	1	1	4	6
GPR	8	5	13	4	2	4	10
IF	25	3	28	24	2	4	30
IGBP	5	3	8	1	0	2	3
INHB	3	2	5	1	1	4	6
INTA	6	4	10	3	1	3	7
INTB	4	3	7	3	2	4	9
NHR	7	3	10	5	4	4	13
NOS	2	2	4	0	0	1	1
PSMB	1	1	2	1	1	4	6
RASR	7	4	11	7	4	4	15
SCN	2	1	3	1	1	3	5
WNT	3	2	5	3	2	4	9
+ Genome Duplication							
Total			145				167

Minimum numbers of genetic events (character changes) required to explain the occurrence of gene family members on human chromosomes 2, 7, 12, and 17, given the phylogenies obtained (see Fig. 1) under the tandem duplication and whole genome duplication models. Genetic events: Dup, tandem duplication; Tra, translocation; Del/Tra, deletion or translocation to another chromosome (i.e., not 2, 7, 12, or 17). Abbreviations of gene family names are as given in Table 1.

dicated that the hypothesis of independent gene duplication and translocation was found to be substantially more parsimonious than that of whole-genome duplication (Table 2). We did not count multiple events of genetic rearrangement within chromosomes that would be required to explain current gene order under the hypothesis of whole-genome duplication. In addition, we assumed that the relationships among genes on chromosomes 2, 7, 12, and 17 are as in Figure 4d because, of all possible phylogenies for these genes, this phylogeny requires fewer translocation events under the hypothesis of whole-genome duplication. However, if this phylogeny truly represented the relationships of genes on the four chromosomes, it would be problematic for the 2R hypothesis, because this phylogeny does not have the form (AB) (CD) expected under that hypothesis (Hughes 1999a). Thus, if our test had been conducted under conditions less favorable to the hypothesis of whole-genome duplication, the results would have been even more strikingly favorable to the hypothesis of independent duplication and translocation.

Our analyses indicate that the occurrence of paralogs belonging to two or more gene families on two or more chromosomes cannot, in the absence of phylogenetic analysis, be taken as evidence that these genes duplicated simultaneously. Even conservation of a linkage relationship for a long period of time is not evidence that the genes involved duplicated simultaneously. For example, a syntenic group including *WNT1*, *WNT10B*, *ARF3*, and *ERBB3* is located both on human chromosome 12 and in the pufferfish *Fugu rubripes* (Gellner and Brenner 1999). *WNT1* duplicated from other *WNT10B* and other *WNT* family members before the divergence of deuterostomes and protostomes (Fig. 1); likewise, *ARF3* duplicated from other *ARF* family members before the divergence of deuterostomes and protostomes (Fig. 1). On the other hand, the *ERBB* family members on the *Hox* chromosomes probably duplicated early in vertebrate history, although not simultaneously with the *Hox* clusters (Fig. 4c). Thus, these genes duplicated independently and then were translocated together at least as early as before the divergence of bony fishes and tetrapods (about 450 Mya), and their linkage has since been conserved in both of these lineages.

Together with other recent results (Hughes 1998a, 1999a), the present analyses indicate that, rather than consisting exclusively of genes duplicated simultaneously in blocks, paralogous groups like those on human chromosomes 2, 7, 12, and 17 often consist of genes that have been duplicated at widely different times and brought together independently during the evolution of the genome. Independent translocation events bringing together paralogs from two or more gene families in two or more independent clusters seem likely to occur with a low probability unless the gene families involved include large numbers of members. For this reason, evidence that the linkage arrangements resulting from such events are found frequently in genomes and have been conserved for long periods of evolutionary time would support the hypothesis that linkage patterns can have adaptive significance (Hughes 1998a, 1999b).

Our results show that the linkage relationships seen on present-day human *Hox*-bearing chromosomes are more easily explained assuming no polyploidization events occurred early in vertebrate history than they are on the 2R hypothesis. Of course, these results in themselves do not "disprove" the 2R hypothesis. However, it is worth recalling that, in science, the null hypothesis is generally the hypothesis of no effect; in this case, the hypothesis that polyploidization did not occur (Hughes 1999a). The 2R hypothesis should be accepted only if there is compelling evidence to reject the null hypothesis, evidence that is certainly not

available at present. Furthermore, the fact that the 2R hypothesis is not well-supported of course has no bearing on other hypotheses involving polyploidization. For example, if two rounds of genome duplication by polyploidization did not occur early in vertebrate history, it is still possible that a single round occurred. In addition, there is some evidence for an independent polyploidization event in teleosts (Amores et al. 1998) and for a relatively recent polyploidization event in the yeast *Saccharomyces cerevisiae* (Wolfe and Shields 1997; Friedman and Hughes 2001).

Moreover, there are good reasons for believing that, even if the 2R hypothesis is correct, the impact of such ancient polyploidization events on present-day genomes is likely to be very small. Wolfe and Shields (1997) estimated the polyploidization event in yeast to have occurred about 100 Mya, and Seoighe and Wolfe (1999) estimated that 16% of the yeast proteome shows effects of this duplication. One hundred Mya is probably an underestimate of the time of polyploidization in yeast, which more likely occurred about 200–300 Mya (Friedman and Hughes 2001). In any event, the yeast example indicates that after polyploidization most duplicate genes are lost relatively quickly. Assuming a comparable rate of gene loss in vertebrates after two polyploidization events that occurred between 528 and 750 Mya, it seems unlikely that more than 4%–9% of the proteome of a present-day vertebrate would show the effects of these events. Our evidence from the human *Hox*-bearing chromosomes is consistent with this prediction. The phylogenetic relationships of paralogs located on these chromosomes indicate that, even if ancient polyploidization events occurred, these events played a very minor role at best in giving rise to current linkage arrangements in vertebrates.

METHODS

Phylogenetic Analyses

Genes from 42 families were included in the analyses (Table 1). These were compared with published phylogenies of *Hox* (Zhang and Nei 1996) and collagen (Bailey et al. 1997) families. Information on chromosomal location of human genes was derived from the Mendelian Inheritance in Man and GenBank databases. Amino acid sequences were aligned using the `CLUSTAL W` program (Thompson et al. 1994). Phylogenetic trees were reconstructed by the maximum parsimony (Swofford 1990) and neighbor-joining (NJ) (Saitou and Nei 1987) methods. Any site at which the alignment postulated a gap in any sequence was not used in the analyses. For the NJ method, three different distances were used: the Poisson-corrected amino acid distance, the γ-corrected amino acid distance, and the uncorrected proportion (p) of amino acid difference (Kumar et al. 1993). Because all methods yielded essentially identical results, only the results of NJ trees based on p are presented here. The NJ method does not assume rate constancy, and this distance is preferable in the case of distantly related sequences, as used in these analyses, because it is expected to have the lowest variance (Nei 1991). The reliability of clustering patterns in trees was tested by bootstrapping (1000 pseudoreplicates) (Felsenstein 1985). To save space, only a summary of the results of the phylogenetic analyses is presented here. All alignments, phylogenetic trees, and accession numbers of sequences used are available from the authors on request.

Duplication Time Estimates

For families and portions of families in which the phylogeny did not indicate duplication before the origin of vertebrates or after the origin of tetrapods, we tested the assumption of the molecular clock. We rejected the hypothesis of a molecular clock if either (1) Takezaki's two-cluster test (Takezaki et al. 1995) using the γ-corrected amino acid distance rejected rate constancy at the 5% level or lower for any pair of branches involved in the comparison between genes on human chromosomes 2, 7, 12, or 17; or (2) in a linearized tree (Takezaki et al. 1995) constructed on the basis of the γ-corrected amino acid distance, organismal divergence points were not proportional to Kumar and Hedges' (1998) estimates of the divergence times of major vertebrate taxa. γ parameters were estimated separately for each family by the maximum likelihood method (Yang 1997). When the molecular clock was not rejected, we estimated gene duplication times from the linearized trees using Kumar and Hedges' (1998) divergence time estimates as calibrations. For the following families having members on just two of the human *Hox*-bearing chromosomes (chromosomal locations in parentheses), the molecular clock hypothesis was rejected, and the phylogeny itself did not provide evidence regarding the duplication time of the genes: acetyl-coA carboxylase (12,17); *NRAMP* (2,12); *even-skipped* (2,7); *frizzled* (2,7); hepatocyte nuclear factor (12,17); myosin light chain (2,17); NAB transcriptional regulator (2,12); pancreatic polypeptide/neuropeptide Y (7,17); peroxidase (2,17).

ACKNOWLEDGMENTS

This research was supported by grant GM34940 to A.L.H. from the National Institutes of Health and by a grant from the South Carolina Commission on Higher Education. We are grateful to F. Verra for comments on the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Amores, A.A., Force, Y.L., Yan, L., Joly, C., Amemiya, A., Fritz, R.K., Ho, J., Langeland, V., Prince, Y.L., Wang, M., et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282:** 1711–1714.

Bailey, W.J., Kim, J., Wagner, G., and Ruddle, F.H. 1997. Phylogenetic reconstruction of the vertebrate *Hox* cluster duplication. *Mol. Biol. Evol.* **14:** 843–853.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 95–105.

Friedman, R. and Hughes, A.L. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11:** 373–381.

Gellner, K. and Brenner, S. 1999. Analysis of 148 kb of genomic DNA around the *wnt* locus of *Fugu rubripes*. *Genome Res.* **9:** 251–258.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B* **256:** 119–124.

———.1998a. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9,

and 1. *Mol. Biol. Evol.* **15:** 854–870.
———. 1998b. Protein phylogenies provide evidence of a radical discontinuity between arthropod and vertebrate immune systems. *Immunogenetics* **47:** 283–296.
———. 1999a. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48:** 565–576.
———. 1999b. *Adaptive evolution of genes and genome*. Oxford University Press, New York.
———. 2001. Evolution of the integrin α and β protein families. *J. Mol. Evol.* **52:** 63–72.
Kasahara, M., Nayaka, J., Satta, Y., and Takahata, N. 1997. Chromosomal duplication and the emergence of the adaptive immune system. *Trends Genet.* **13:** 90–92.
Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.
Kumar, S., Tamura, K., and Nei, M. 1993. *MEGA: Molecular Evolutionary Genetics Analysis, Version 1.01*. Pennsylvania State University, University Park, PA.
Lundin, L.G. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosome regions in man and the house mouse. *Genomics* **16:** 1–19.
Miklos, G.L.G. and Rubin, G.M. 1996. The role of the genome project in determining gene function: Insight from model organisms. *Cell* **86:** 521–529.
Nei, M. 1991. Relative efficiencies of different tree-making methods for molecular data. In *Phylogenetic analysis of DNA sequences* (eds. M.M. Miyamoto and J.L. Cracraft), pp. 90–128. Oxford University Press, Oxford.
Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Ohno, S. 1970. *Evolution by gene duplication.* Springer Verlag, New York.
Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
Seoighe, C. and Wolfe, K.H. 1999. Updated map of duplicated regions in the yeast genome. *Gene* **238:** 253–261.
Sidow, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6:** 715–722.
Simmen, M.W., Leitgeb, S., Clark, V.H., Jones, S.J.M., and Bird, A. 1998. Gene number in an invertebrate chordate, *Ciona intestinalis*. *Proc. Natl. Acad. Sci.* **95:** 4437–4440.
Skrabanek, L. and Wolfe, K.H. 1998. Eukaryotic gene duplication—where's the evidence? *Curr. Opin. Genet. Dev.* **8:** 694–700.
Swofford, D.L. 1990. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0*. Illinois Natural History Survey, Champaign, IL.
Takezaki, N., Rzhetsky, A., and Nei, M. 1995. Phylogenetic test of the molecular clock and linearized tree. *Mol. Bio. Evol.* **12:** 823–833.
Thompson, J.D., Higgins, D.G., and Gibson, J.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Wolfe, K.H. and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.
Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yeager, M. and Hughes, A.L. 1999. Evolution of the mammalian MHC: Natural selection, recombination, and convergent evolution. *Immunol. Rev.* **167:** 45–58.
Zhang, J. and Nei, M. 1996. Evolution of *Antennapedia*-class homeobox genes. *Genetics* **142:** 295–303.

Received August 10, 2000; accepted in revised form February 14, 2001.