Homologs of *Drosophila P* Transposons Were Mobile in Zebrafish but Have Been Domesticated in a Common Ancestor of Chicken and Human

Sabine E. Hammer, Sabine Strehl,† and Sylvia Hagemann**

*Laboratories of Genome Dynamics, Center of Anatomy and Cell Biology, Medical University of Vienna, Vienna, Austria; and †Children's Cancer Research Institute, Vienna, Austria

A substantial fraction of vertebrate and invertebrate genomes is composed of mobile elements and their derivatives. One of the most intensively studied transposon families, the *P* elements of *Drosophila*, was thought to exist exclusively in the genomes of dipteran insects. Based on the data provided by the human genome project, in 2001 our group has identified a *P* element–homologous sequence in the human genome. This *P* element–homologous human gene, named *Phsa*, is 19,533 nucleotides long, comprises six exons and five introns, and encodes a protein of still unknown function with a length of 903 amino acid residues. The N-terminal THAP domain of the putative Phsa protein shows similarities to the site-specific DNA-binding domain of the *Drosophila P* element transposase. In the present study, FISH analysis and the screening of a human lambda genomic library revealed a single copy of *Phsa* located on the long arm of chromosome 4, upstream of a gene coding for the hypothetical protein DKFZp686L1814. The same gene arrangement was found for the homologous gene *Pgga* in the genome of chicken, thus, displaying *Pgga* at orthologous position on the long arm of chromosome 4. The single-copy gene status and the absence of terminal inverted repeats and target-site duplications indicate that *Phsa* and *Pgga* constitute domesticated stationary sequences. In contrast, a considerable number of *P*-homologous sequences with terminal inverted repeats and intact target-site duplications could be identified in zebrafish, strongly indicating that *Pdre* elements were mobile within the zebrafish genome. *Pdre* elements are the first *P*-like transposons identified in a vertebrate species. With respect to *Phsa*, gene expression studies showed that *Phsa* is expressed in a broad range of human tissues, suggesting that the putative Phsa protein plays a not yet understood but essential role in a specific metabolic pathway. We demonstrate that *P*-homologous DNA sequences occur in the genomes of 21 analyzed vertebrates but only as rudiments in the rodents. Finally, the evolutionary history of *P* element–homologous vertebrate sequences is discussed in the context of the "molecular domestication" hypothesis versus the "source gene hypothesis."

Introduction

The *P* elements of *Drosophila* belong to the class of DNA transposons. Autonomous elements are about 3 kb long and consist of four exons, three introns, and terminal noncoding sequences ending in terminal inverted repeats (O'Hare and Rubin 1983). Their transposition follows a cut-and-paste mechanism, generating an 8-bp target-site duplication. The transposition event is catalyzed by the 87-kDa transposase that is translated in germline cells from a mRNA transcribed from the four exons. In somatic cells, a 66-kDa protein is synthesized from an mRNA that retains the third intron, leading to a premature stop of translation. This truncated protein acts as a repressor of *P* element transposition (Misra and Rio 1990). *P* elements were first discovered in *Drosophila melanogaster* because of their ability to induce hybrid dysgenesis (Kidwell, Kidwell, and Sved 1977). Subsequently, *P* element–homologous sequences were identified in many drosophilid species (for review, see Pinsker et al. [2001]) and in some dipteran species outside the drosophilid family (Perkins and Howells 1992; Lee, Clark, and Kidwell 1999; Sarkar et al. 2003; Oliveira de Carvalho, Silva, and Loreto 2004). Sequence analyses of different *P* element subfamilies within the Drosophilidae revealed that their sequence relationships are not in accordance with the phylogeny of their host species (Hagemann, Haring, and Pinsker 1996; Haring, Hagemann, and Pinsker 2000; Silva and Kidwell 2000). These findings resulted in the generally accepted model that *P* elements are not only vertically inherited but also can be transmitted horizontally. The first horizontal transmission identified was the transfer from *D. willistoni* to *D. melanogaster* (Daniels et al. 1990). This rather recent event was followed by the rapid spread of *P* elements through the natural populations of *D. melanogaster*. A considerable number of horizontal transfer events must be considered to explain the present distribution pattern of *P* element subfamilies within the Drosophilidae (Pinsker et al. 2001; Silva and Kidwell 2000). In some *Drosophila* species (Miller et al. 1992; Nouaud and Anxolabéhère 1997), in the blowfly *Lucilia cuprina* (Perkins and Howells 1992), and in the housefly *Musca domestica* (Lee, Clark, and Kidwell 1999) terminally truncated and, therefore, immobile *P* transposons have been detected. In *Drosophila*, these stationary sequences have retained the coding capacity of the first three exons, thus, expressing a repressor-like protein with unknown function. This change from a parasitic element to a beneficial host gene has been described as molecular domestication (Miller et al. 1999).

In humans, at least 45% of the genome belong to transposable elements, and a number of single-copy genes seem to have originated from them. Until now, 48 domesticated human genes probably originating from up to 39 different transposon copies could be identified (Hagemann and Pinsker 2001; International Human Genome Sequencing Consortium 2001; Nekrutenko and Li 2001). Most of them originated from DNA transposons, although only about 6% of the human transposable elements belong to this transposon type.

In 2001 we reported the detection of a *P*-homologous sequence in the human genome (Hagemann and Pinsker 2001), which we subsequently named *Phsa* (*P* homolog of *Homo sapiens*). From the sequences available at that time, we concluded that *Phsa* codes for a protein of 759 amino

Key words: *Drosophila P* transposons, *Homo sapiens*, *Danio rerio*, *Gallus gallus*, THAP proteins, molecular domestication.

E-mail: sylvia.hagemann@meduniwien.ac.at.

Mol. Biol. Evol. 22(4):833–844. 2005
doi:10.1093/molbev/msi068
Advance Access publication December 22, 2004

Table 1
PCR Primers Used in This Study

Primer Name	Sequence 5′ → 3′	Location	Position
Pver1+	GGCTCAGCTGCTTCGTC	*Phsa* exon 5	1050[a]
Pver1−	GCTCTCAGCATTGAGCAAA	*Phsa* exon 5	1749[a]
q1+Phsa	GTGTTCAGATGGCAAAAGCA	*Phsa* exon 5	1136[a]
q1−Phsa	GGGTGGCACTATTCACTTTCA	*Phsa* exon 5	1438[a]
q1+G3PDH	AATCCCATCACCATCTTCCA	Accession number BT006893	
q1−G3PDH	TGTGGTCATGAGTCCTTCCA	Accession number BT006893	
Pgga+3	GGAGGGGGAATTGCACATTGACAATCA	*Pgga* exon 2	1839[b]
Pgga+4	CCACTGCAGCAGGCACAAATTGAAGAG	*Pgga* exon 2	1947[b]

[a] Position of the 5′ end of the primers are relative to the ATG start codon of the human *Phsa* cDNA (accession number AK091412).
[b] Position of the 5′ end of the primers are relative to the chicken *Pgga* cDNA (accession number XM_420555).

acids. Closer examination of the genomic organization of *Phsa* as well as the comparison with several insect *P* element sequences revealed the absence of transposon characteristic terminal inverted repeats and led to the assumption that *Phsa* is a stationary sequence in the human genome (Hagemann and Pinsker 2001). Through database searches, *P*-homologous sequences have been identified also in the genomes of chicken (*Gallus gallus*) and cattle (*Bos taurus*) (Hagemann and Pinsker 2001), indicating a widespread occurrence in vertebrates.

Roussigne et al. (2003*a*) described a novel proapoptotic factor, which they designated THAP1. Database searches with both the nucleotide and the amino acid sequences of human THAP1 revealed that the first 89 N-terminal amino acid (aa) residues constitute a novel protein motif, the THAP domain, which seems to be evolutionary conserved and restricted to animals. They suggested that THAP domain proteins might belong to a new family of cellular DNA-binding proteins (Roussigne et al. 2003*b*) and demonstrated that the THAP domain shows striking similarities to the site-specific DNA-binding domain of the *D. melanogaster P* element transposase. They designated this *P* element transposase–homologous protein THAP9 (THAP domain containing 9), which corresponds to the putative *Phsa* gene product detected in 2001 (Hagemann and Pinsker 2001) and further characterized in this paper.

In the present study, we prove experimentally that the human *Phsa* is a single-copy gene located on the long arm of chromosome 4. *Pgga* (*P* homolog of *Gallus gallus*), the homologous gene from chicken, as well as a mouse and a rat *P*-homologous rudiment, can be found at orthologous positions. With respect to zebrafish, Blast searches resulted in the detection of a considerable number of *P*-homologous sequences (*Pdre* [*P* homolog of *Danio rerio*]) with DNA-transposon characteristic structural features such as terminal inverted repeats and target-site duplications, strongly indicating that they were mobile within the zebrafish genome. To gain some insights into the possible function of the domesticated *Phsa*, we carried out expression studies to characterize the transcription pattern in different human tissues. Finally, we describe the phylogenetic relationship of *Phsa*-homologous sequences of various vertebrates and discuss their evolutionary history as active transposable elements versus domesticated genes.

Materials and Methods

Primer Design to Create *P*-Homologous Probes

By aligning the cDNA sequences of the *P*-homologous elements from *Homo sapiens* (accession number NM_024672), *Bos taurus* (accession number AW483725), and *Gallus gallus* (accession numbers AJ395159 and AJ394151), the most conserved region in the vertebrate *P* homologs was determined. Based on this sequence alignment, primer pair Pver1+ and Pver1− (table 1), which is located in exon 5 of the *Phsa* gene and amplifies a 700-bp fragment, was designed (figs. 1*A* and 4*B*).

Screening of a Human Genomic Library

A lambda DASH II library (Stratagene, La Jolla, Calif.) was screened for clones containing human *P*-homologous sequences using a 700-bp digoxigenin-labeled (DIG-labeled) probe that was PCR amplified from 10 ng of human genomic DNA (EMD Biosciences, San Diego, Calif.) with the primers Pver1+ and Pver1− (table 1). This fragment will be referred to as *Phsa*700-G (fig. 1*A*). PCR was carried out on an Eppendorf Mastercycler Gradient in 50-μl reactions containing 2.5 μM $MgCl_2$, each dNTP at 200 μM, each primer at 50 pM, 2.5 units of *Taq* polymerase (Promega, Madison, Wis.), and the reaction buffer supplied by the manufacturer. After an initial denaturation step of 2 min at 94°C, 35 PCR cycles with 15 s at 94°C denaturation, 15 s at 53°C annealing, and 45 s at 74°C extension were performed, followed by a final extension step of 10 min at 74°C. The titering procedure of the lambda library and the plaque lifts were performed according to standard protocols (Ausubel et al. 1998). Hybridization and washing were performed under the following stringency conditions: hybridization overnight in 5 × SSC/0.02% SDS at 68°C, washing twice for 5 min in 2 × SSC/1% SDS at room temperature, and washing twice for 15 min in 0.1% SSC/0.1% SDS at 68°C. The identification of positive plaques was carried out using the DIG detection system (Roche Diagnostics, Mannheim, Germany) following the manufacturer's instructions.

Fluorescence in situ Hybridization (FISH) Analysis

To identify the location of the *Phsa* gene and its possible homologs on human chromosomes, FISH using a

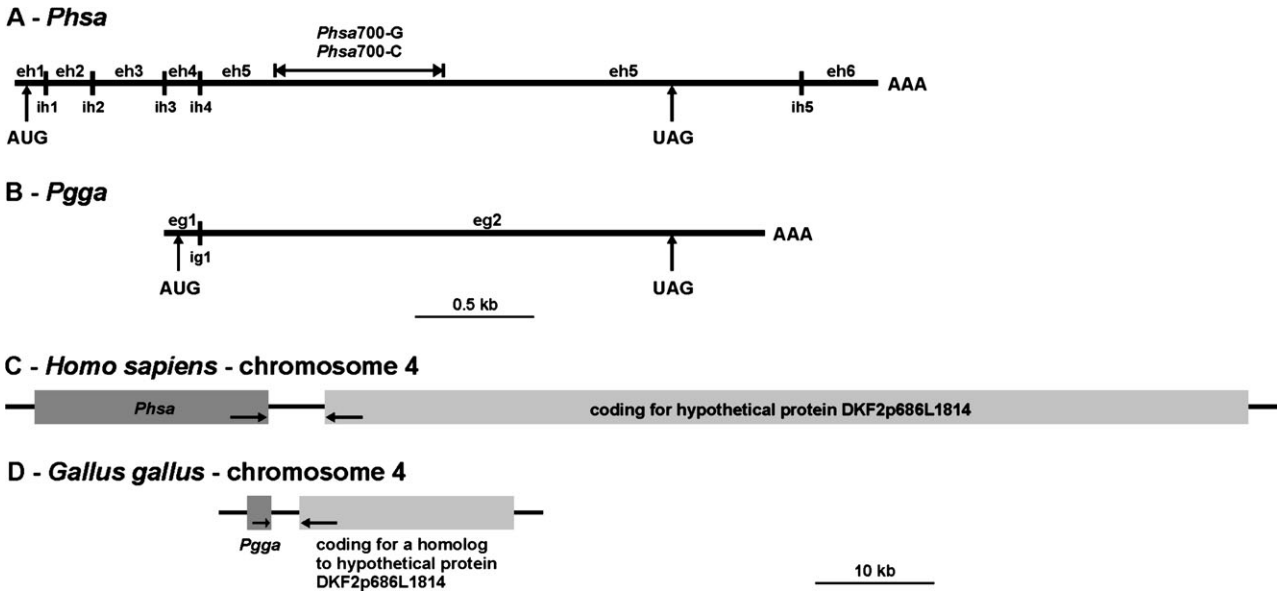


FIG. 1.—Schematic representation of the human *Phsa* (*A*) and the chicken *Pgga* (*B*) gene transcript, as well as the orthologous position of *Phsa* (*C*) and *Pgga* (*D*), on chromosome 4 in the respective genomes. The thick line represents the mRNA: …mRNA(A,B) and the genomic DNA (C,D). Vertical bars indicate the exon/intron borders. The double-headed arrow represents the probe for subsequent hybridization experiments. Exons are designated as eh (*Phsa*) and eg (*Pgga*); introns are designated as ih (*Phsa*) and ig (*Pgga*). Horizontal arrows indicate the orientation of transcription.

locus-specific probe was performed. For this purpose, a 15-kb lambda phage clone covering the 3′ end of *Phsa* that was obtained by screening a human genomic library was used. The probe was labeled with DIG-II-dUTP by nick translation, and FISH was carried out as previously described in Koenig et al. (2002). For immunodetection of the probe, sheep antidigoxigenin FITC (1:100) (Roche Diagnostics, Mannheim, Germany) and rabbit anti–sheep FITC (1:100) (Dako, Hamburg, Germany) were used.

Genome-Walking Analysis in Chicken

To determine the copy number of *P*-homologous sequences within the genome of chicken, genome-walking experiments were carried out using the Universal Genome Walker Kit of CLONTECH Laboratories (BD Biosciences, San Jose, Calif.). Genomic DNA was digested with the restriction enzymes *Dra*I, *Eco*RV, *Pvu*II, and *Stu*I and genome walking was performed towards the 3′ end of the target gene following the manufacturer's instructions. The nested primers Pgga+3 and Pgga+4 served as gene-specific primers, binding in exon 2 of the *Pgga* gene (table 1 and fig. 1*B*). The obtained PCR products were cloned using the pGEM-T Easy vector system and subjected to automated sequencing.

Quantitative Real-Time PCR

To examine the levels of specific transcripts of *Phsa* in different human tissues, multiple tissue cDNA panels (Human MTC Panel [BD Biosciences, San Jose, Calif.]) were analyzed. These cDNA panels contain normalized, first-strand cDNA preparations from RNA of the following pooled human tissues or cell lines that were derived from at least two male and/or female Caucasians: brain (whole), heart, kidney, liver, lung, pancreas, peripheral blood leukocyte, placenta, and skeletal muscle (for details regarding the normalization procedure, see the cDNA Panels User Manual). The human cDNA samples were subjected to quantitative real-time PCR (qPCR). A 303-bp fragment of exon 5 of *Phsa* was amplified with the primers q1+Phsa and q1−Phsa (table 1) using the LightCycler DNA Master SYBR Green I kit (Roche Diagnostics, Mannheim, Germany). In this assay, the housekeeping gene (HK) for glycerol-3-phosphate dehydrogenase (G3PDH, EC 1.1.1.8) served as positive control, and the respective qPCR was conducted with the primer pair q1+G3PDH and q1−G3PDH (table 1), which lead to a 318-bp PCR product. The qPCR was performed on the LightCycler instrument (Roche Diagnostics, Mannheim, Germany) in 15-µl reactions containing 1.5 µl FastStart DNA Master reaction mix, 0.65 µM $MgCl_2$, 5 pM of each primer, and 1 ng/µl normalized cDNA sample. The PCR conditions were 95°C for 10 min, followed by 45 cycles at 95°C for 10 s, 59°C (*Phsa* and *G3PDH*) for 5 s, and 72°C for 20 s (slopes were 20°C/s). Fluorescence was measured at the end of the extension phase. To confirm the specificity of the amplified products, melting curves were performed at the end of the amplification by cooling the samples at 20°C/s to the respective annealing temperatures of both primer pairs and then increasing the temperature to 95°C at 0.2°C/s with fluorescence measurement every 0.1°C. Two standard curves for *Phsa* and *G3PDH*, respectively, were generated using serial dilutions of the liver sample of the cDNA panel at concentrations of 0.008, 0.01, 0.013, 0.02, 0.05, 0.1, 0.2, and 1 ng/µl. All standards were amplified in duplicates and a regression curve was computed, which served as internal reference for further calculations. The identical cDNA samples were examined at four different times, and in each experiment, the samples were

Table 2
Vertebrate Specimens Examined in the Phylogenetic Analysis

Scientific Name (Common Name)	Systematic Position	*P*-Homologous Fragment	DNA Sequence Homology to the Human *Phsa* (%)
Talpa europaea (European mole)	Insectivora	Yes	92.56
Sorex alpinus (Alpine shrew)	Insectivora	Yes	83.70
Plecotus auritus (brown long-eared bat)	Insectivora	Yes	89.84
Lepus europaeus (European brown hare)	Lagomorpha	Yes	87.70
Oryctolagus cuniculus (European rabbit)	Lagomorpha	Yes	87.27
Mus musculus (house mouse)	Rodentia	No	—
Rattus rattus (black rat)	Rodentia	No	—
Rattus norvegicus (brown rat)	Rodentia	No	—
Microtus oeconomus (root vole)	Rodentia	No	—
Castor fiber (European beaver)	Rodentia	No	—
Felis catus (domestic cat)	Carnivora	Yes	89.14
Canis familiaris (domestic dog)	Carnivora	Yes	89.57
Mustela putoris (ferret)	Carnivora	Yes	83.70
Equus caballus (domestic horse)	Perissodactyla	Yes	92.70
Bos taurus (cattle)	Artiodactyla	Yes	87.98
Lemur catta (ring-tailed lemur)	Primates	Yes	79.11
Saimiri sciureus (squirrel monkey)	Primates	Yes	97.00
Macaca mulatta (rhesus monkey)	Primates	Yes	98.27
Hylobates lar (gibbon)	Primates	Yes	99.00
Pongo pygmaeus (orangutan)	Primates	Yes	99.14
Pan paniscus (pygmy chimp, bonobo)	Primates	Yes	99.57
Pan troglodytes (chimpanzee)	Primates	Yes	99.43
Gorilla gorilla (gorilla)	Primates	Yes	99.43
Homo sapiens (human)	Primates	Yes	100.00
Gallus gallus (chicken)	Aves (Neornithes)	Yes	58.94
Danio rerio (zebrafish)[a]	Teleostei (Ostariophysi)	Yes	50.50

[a] The *P*-homologous zebrafish sequence (*Pdre2*) was taken from GenBank (accession number BX511023).

measured in duplicates. For the statistical analysis, means, standard deviations, and standard errors were computed. The relative abundance of *Phsa* gene transcript was measured by calculating the ratio of mean concentrations of this gene over the mean concentrations of *G3PDH* for the respective tissue samples. Differences between *Phsa* and *G3PDH* synthesis were evaluated by one-way analysis of variance (1-way ANOVA), followed by the paired samples (two sample) *t*-test, respectively, both as implemented in Analyse-it version 1.71 (Analyse-it Software Ltd.) for Microsoft Excel.

Phylogenetic Analysis of *P*-Homologous Sequences

Details regarding the 26 vertebrate species analyzed in this study are given in table 2. Total genomic DNA of the eight primates was provided by the German Primate Center (Göttingen, Germany), commercially available DNA was used for an additional eight vertebrates (EMD Biosciences, San Diego, Calif.), and the DNA of the remaining nine specimens was extracted following the protocol described in the Genetic Analysis Manual (LI-COR, Inc. 1999). Depending on the quality and amount of genomic DNA, 1 to 5 μl (≤1 μg) were subjected to PCR amplification using primers Pver1+ and Pver1− (table 1). PCR reactions were electrophoresed on 1% agarose/EtBr gels and blotted onto nylon membranes (Amersham Biosciences, Uppsala, Sweden) by Southern transfer following standard protocols (Sambrook, Fritsch, and Maniatis 1989). For hybridization, a 700-bp fragment of exon 5 of *Phsa* was amplified with the primers Pver1+ and Pver1− and cloned using the pGEM-T Easy vector system (Promega, Madison, Wis.). This probe will be referred to as *Phsa*700-C (fig. 1*A*). After PCR DIG labeling, this probe was used to identify homologs of *Phsa* with the DIG detection system. Positive PCR fragments were cloned using the pGEM-T Easy vector system and subjected to automated sequencing. *P*-homologous DNA sequences (719 bp) and the deduced amino acid sequences (239 aa) of 21 vertebrate species (table 2) were aligned by the program ClustalX (Thompson et al. 1997). A maximum-likelihood (ML) analysis was conducted by application of the quartet-puzzling approach of Tree-Puzzle (Schmidt et al. 2002). The substitution process was modeled by implementing the HKY85 model (Hasegawa, Kishino, and Yano 1985) for the DNA sequence data, and the Dayhoff model (Dayhoff, Schwartz, and Orcutt 1978) for the protein data, respectively. The DNADIST option of PHYLIP (Felsenstein 2004) was used to perform a Neighbor-Joining analysis (NJ [Saitou and Nei 1987]) based on the F84 model (Felsenstein and Churchill 1996) of nucleotide evolution and the JTT92 model (Jones, Taylor, and Thornton 1992) of amino acid sequence evolution. To evaluate the quality of the phylogenetic signal of *Phsa* for resolving phylogenetic relationships among vertebrates, the obtained *Phsa* gene trees were compared with those of the mitochondrial (mt) cytochrome *b* (*cyt b*). Therefore, the entire mtDNA *cyt b* gene DNA sequences (1,149 bp) and the deduced amino acid sequences (382 aa) of 25 vertebrate species (table 2) were aligned with the program ClustalX (for accession numbers, see *Supplementary Material*). The mtREV24 model (Adachi and Hasegawa 1996) and the JTT92 model of amino acid sequence evolution, as

well as the TrN model (Tamura and Nei 1993) and the F84 model of nucleotide evolution, were used for likelihood and distance analysis. The robustness of the phylogenies was assessed by the bootstrap percentage, and by the reliability percentages, (i.e., the number of times the group appears after 10,000 puzzling steps).

Results and Discussion

Genomic Organization of *Phsa* in *Homo sapiens*

The *Phsa* gene has a total length of 19,533 nucleotides and consists of six exons and five introns (fig. 1*A*). The region of interest extends from nt 149213 to nt 129681 of BAC-clone RP11-163O17 (accession number AC021105 [Sulston and Waterston 1998]). *Phsa* is transcribed into a 3,627-bp transcript (accession number NM_024672), which codes for a hypothetical protein of 903 aa with unknown function (accession number NP_078948). To determine the copy number of *Phsa*, library screening and FISH analyses were performed. The screening of a human lambda DASH II genomic library revealed only one single *Phsa*-positive lambda clone (total length of insert: 15 kb), which carried exons 5 and 6, thus encompassing the 3′ end of the *Phsa* gene (fig. 1*A*). The 4,978-bp *Phsa*-specific part of the lambda clone matched nt 133705 to nt 128727 of BAC-clone RP11-163O17 (accession number AC021105 [Sulston and Waterston 1998]) localized at chromosome region 4q21.3. Using the *Phsa*-containing lambda clone obtained by library screening as a probe for FISH, only one single hybridization signal was observed (fig. 2*A* and *B*). This single copy of *Phsa* was located on the long arm of chromosome 4, thus confirming the result of the lambda library screening as well as previously conducted Blast searches (Hagemann and Pinsker 2001). The data obtained by the genomic library screening and the FISH analysis provide compelling evidence that *Phsa* is a single-copy gene in the human genome.

Further analyses of the BAC-clone RP11-163017 (accession number AC021105) led to the detection of a neighboring gene, which is transcribed into the cDNA with accession number BX640657 and codes for the hypothetical protein DKFZp686L1814 (fig. 1*C*). This gene extends from nt 39137 to nt 124991 and was used to provide evidence for the orthologous positions of the human *Phsa* and the chicken *Pgga* as described in the next section.

Genomic Organization of the *P*-Homologous *Pgga* in *Gallus gallus*

In 2001, we reported the discovery of a *P*-homologous sequence in chicken (Hagemann and Pinsker 2001), which will be referred to as *Pgga*. The genome-walking experiments revealed a single copy of *Pgga*, which corresponds to the sequence represented by accession number NW_060347, confirming the Blast search results suggesting that *Pgga* is a single-copy gene like its human homolog, *Phsa*. The *P*-homologous region extends from nt 16680605 to nt 16676896, comprises a region of 3,710 bp, and is located on the long arm of chromosome 4. Based on the alignment of the cDNA sequence (accession number XM_420555) with the corresponding genomic sequence, we were able to determine the exon/intron limits of this gene. *Pgga* consists of one intron (ig = 1,185 bp) and two exons (eg1 = 156 bp, eg2 = 2,369 bp [fig. 1*B*]), which can be translated into a *P* element–homologous protein of 681 amino acid residues. The intron belongs to the minor group of GC–AG introns, the 5′ splice donors of which are thought to play an important factor in the regulation of alternative splicing (Farrer et al. 2002). Exon eg1 is homologous to the human exon eh4, whereas eg2 is homologous to the human eh5 (fig. 1*A* and *B*), and the start codon of *Pgga* corresponds to the previously assumed start codon of the first human cDNA entry (accession number AK026973).

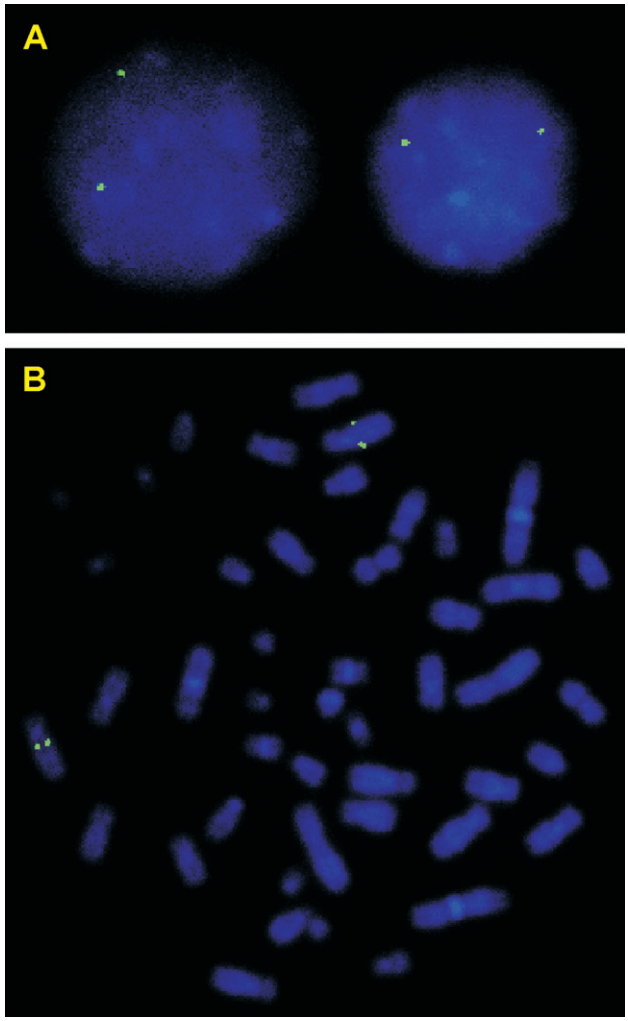


FIG. 2.—Fluorescence in situ hybridization (FISH) analysis using the *Phsa*-specific lambda clone. (*A*) Interphase nuclei and (*B*) metaphase chromosomes displaying only one single hybridization signal indicating the localization of *Phsa* on the long arm of chromosome 4 (4q21).

In comparison with the putative Phsa protein, the presumed *Pgga* gene product shows a N-terminal truncation, thus, lacking the DNA-binding THAP domain (fig. 1*A* and *B* and fig. 4*A* and *B*). Downstream of *Pgga*, the functional gene LOC422595 (cDNA accession number XM_420554) was identified, giving rise to a protein similar to the hypothetical protein DKFZp686L1814 located downstream of *Phsa* in human (fig. 1*D*). These findings strongly support

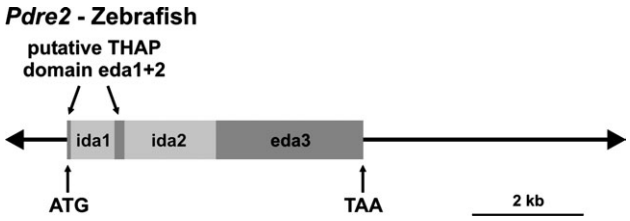


FIG. 3.—Schematic representation of the *P* element homolog *Pdre2* in zebrafish (accession number BX511023). The thick line represents the genomic DNA (11,149 bp). The dark gray and light gray boxes indicate the putative exons (eda) and the putative introns (ida), respectively. Arrows indicate the terminal inverted repeats.

the orthologous positions of *Phsa* and *Pgga* in their respective host genomes (fig. 1*C* and *D*).

Genomic Organization of *P*-Homologous *Pdre* Elements in *Danio rerio*

Blast searches with the deduced amino acid sequence of *Phsa* as query sequence revealed four different Blast hits in the zebrafish genome. The homologous sequences are located in linkage group 2 (accession number BX511023), group 4 (accession number BX890548), group 17 (accession number BX324003), and group 22 (accession number CR388079) and will be referred to as *Pdre2*, *Pdre4*, *Pdre17*, and *Pdre22*. Multiple alignment analyses of these sequences showed that the repeated region of *Pdre2* in linkage group 2 is the most extended one. Consequently, this sequence was used for further Blast searches, leading to the detection of at least 50 *P*-homologous sequences spread throughout the genome of zebrafish. Most of them are internally deleted copies (e.g., in linkage groups 1 [accession number BX537346: nt 47346 to nt 48360], 3 [accession number AL929078: nt 95952 to nt 96777], and 20 [accession number AL954815: nt 148549 to nt 149356]) referred to as *Pdre1*, *Pdre3*, and *Pdre20*, whereas *Pdre4* (accession number BX890548: nt 39502 to nt 47547), *Pdre17* (accession number BX324003: nt 50600 to nt 56639), and *Pdre22* (accession number CR388079: nt 17390 to nt 27566) are terminally truncated. One putative complete transposon was found in linkage group 2 (*Pdre2* [fig. 3]), which is 11,149 bp long (accession number BX511023: nt 198423 to nt 209571), possesses 13-bp terminal inverted repeats (5′-CATACCTGTCAAC-3′/5′-GTTGACAGG-TATG-3′) together with 12-bp subterminal inverted repeats (5′-TGTTTAAACCAA-3′/5′-TTGGTTTAAACA-3′), and is flanked by an 8-bp target-site duplication (5′-AGGT-GAAT-3′). Terminal inverted repeats as well as the *Drosophila P* transposon–typical 8-bp target site duplications can also be found for the analyzed internally deleted elements *Pdre1* (5′-CCTTTAAG-3′), *Pdre3* (5′-CATC-CAAC-3′), and *Pdre20* (5′-GTCTACAT-3′), but they are missing for the terminally truncated elements *Pdre4*, *Pdre17*, and *Pdre22*. The large sizes of *Pdre2* (11,149 bp) and of the terminally truncated elements *Pdre4* (8,046 bp), *Pdre17* (5,940 bp), and *Pdre22* (10,177 bp) indicate that this is a conserved feature of *Pdre* elements.

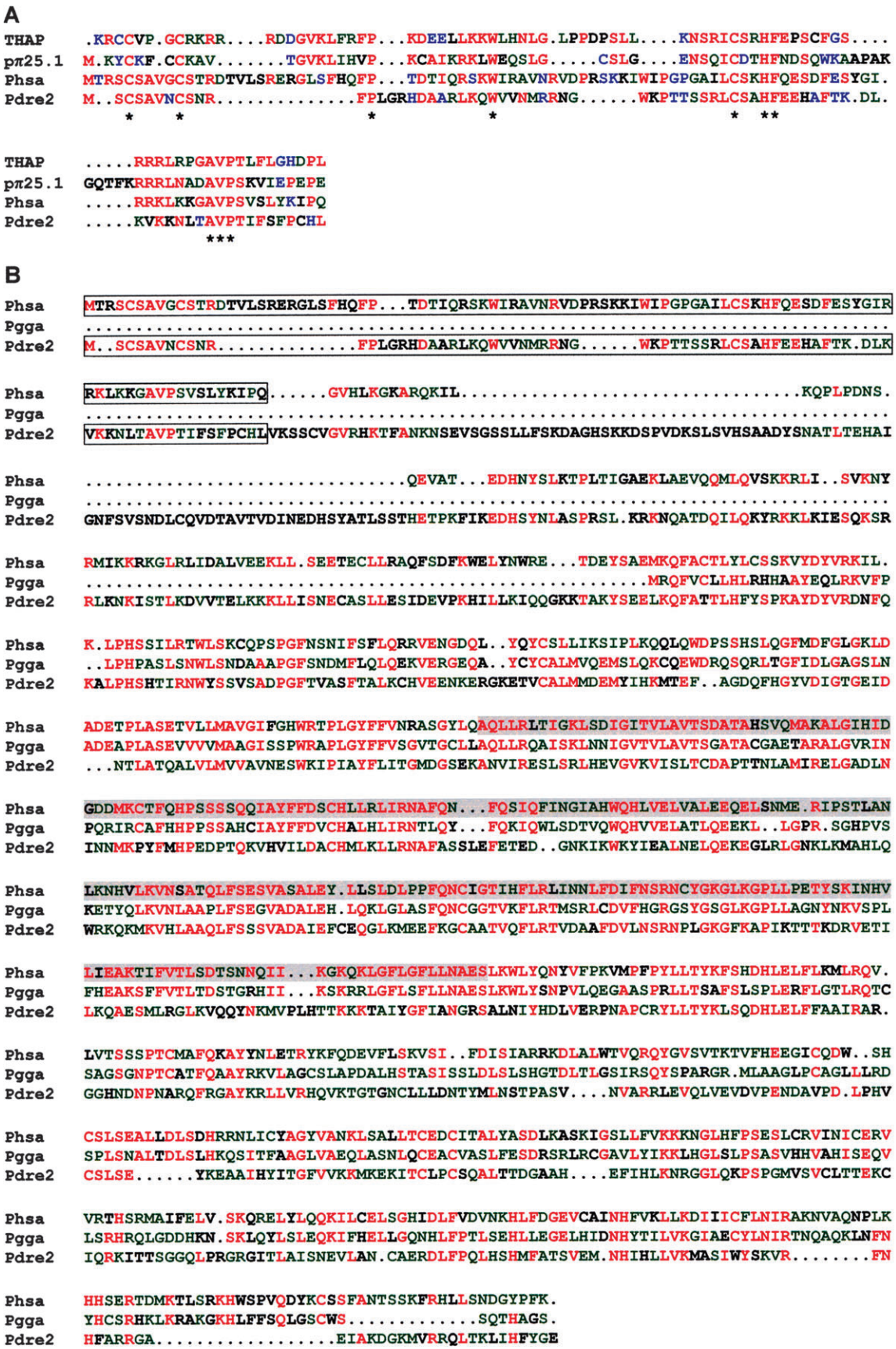
Comparing the terminal inverted repeats of the *Pdre* elements with those of different *P* element families from insects revealed that only the first (5′-CAT-3′) and the last three (5′-ATG-3′) bases are conserved. In contrast to the *Drosophila P* element transposons, where different *P* element families with DNA sequence divergences of about 30% can be found in some species (for review, see Pinsker et al. [2001]), all analyzed *Pdre* elements belong to the same family. Comparing the homologous internal regions of *Pdre2*, *Pdre4*, *Pdre17*, and *Pdre22* revealed a maximum sequence divergence of 3.5% over 7,603 bp, whereas the terminal regions of *Pdre1*, *Pdre2*, *Pdre3*, and *Pdre20* show sequence divergences between 6.5% and 8.5 % over 452 bp.

With respect to *Pdre2*, the only long open reading frame (ORF) extends from nt 3448 to nt 6078 comprising a region of 2,631 bp that can be translated into a *P* element–homologous protein of 877 amino acid residues. Because a 5′ start codon is lacking within this ORF, we performed translation product analyses of the upstream region by searching for conserved protein domains (Marchler-Bauer et al. 2003). Using the deduced amino acid sequence from the first reading frame as query sequence revealed a THAP domain (E-value: 2×10^{-04}), the putative DNA-binding domain as described by Roussigne et al. (2003*b*). The THAP domain–corresponding DNA sequence comprises the *Pdre2* region from nt 1639 to nt 1806, but the N-terminal region of the THAP domain was missing. Therefore, the upstream region was searched for additional exons and a further small exon could be detected comprising the *Pdre2* sequence from nt 849 to nt 880, suggesting a transposase gene consisting of three exons (eda1 = nt 849 to nt 880, eda2 = nt 1656 to nt 1808, and eda3 = nt 3468 to 6078) and two introns (ida1 and ida2 [fig. 3]). Both introns belong to the type of GC–AG introns indicating that the *Pdre2* gene could be alternatively spliced (Farrer et al. 2002). Based on these findings, we constructed a hypothetical mRNA, which could be translated into a protein of 932 amino acid residues with a N-terminal THAP domain (E-value: 7×10^{-08}) overlapping the eda1/eda2/eda3 boundaries; hence, it follows that our splicing interpretation is correct.

In contrast to *Pdre2*, the ORFs of the putative exon eda3 of the terminally truncated elements *Pdre4*, *Pdre17*, and *Pdre22* are destroyed. Thus, within the zebrafish genome, two scenarios can be observed for the *P*-homologous sequences: *Pdre4*, *Pdre17*, and *Pdre22* are terminally truncated and, therefore, present immobile components of the zebrafish genome, whereas *Pdre2* and the analyzed internally deleted copies *Pdre1*, *Pdre3*, and *Pdre20* have terminal inverted repeats and are flanked by target-site duplications, both typical characteristics for mobile DNA transposons.

Protein Domains of Vertebrate *P*-Homologous Sequences

Roussigne et al. (2003*a*) described a novel protein motif, the THAP domain, which shows striking similarities to the DNA-binding site of *Drosophila P* element transposases. To date, 113 THAP domain–containing proteins are listed in the databases, including seven murine and 12 human THAP proteins, and two of them were known to have metabolic functions (Gale et al. 1998; Smit 1999; Roussigne et al. 2003*a*). The THAP motif contains a putative DNA-binding domain that is characterized by



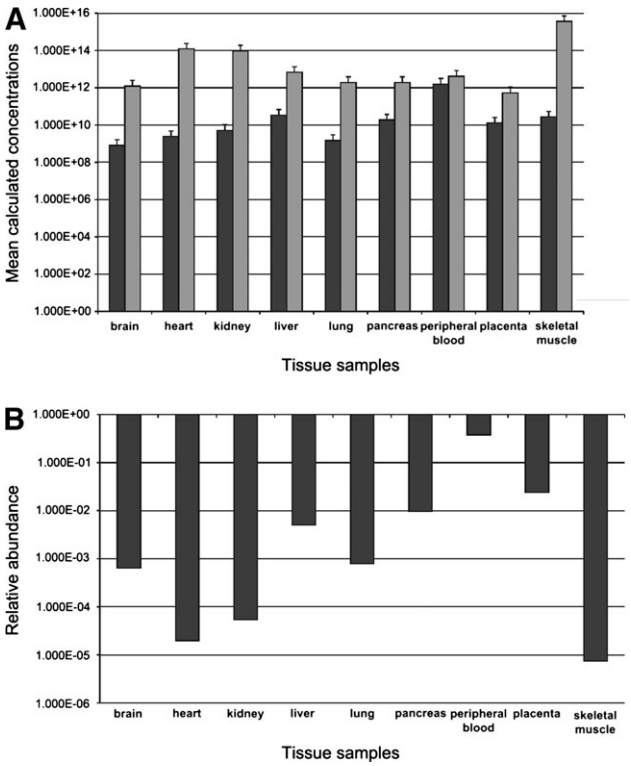


FIG. 5.—Real-time PCR analysis of *Phsa* mRNA in the human multiple tissue cDNA panels. (*A*) The means of calculated concentrations (plus standard deviations) of *Phsa* mRNA and the HK gene *G3PDH* are shown for each tissue analyzed. Expression of *Phsa* (*G3PDH*) in peripheral blood leukocyte (skeletal muscle) differed significantly from the other tissues ($P < 0.05$). (*B*) The relative abundance of *Phsa* mRNA is given for each tissue. *Phsa*-gene expression levels in the tissues analyzed are significantly different (table 3). *Y*-axis is represented as a logarithmic scale.

tetrahedral formation of one or two zinc ions by conserved cysteine and histidine residues. Zinc-coordinating proteins constitute the largest group of transcription factors in eukaryotic genomes and were also found in domains that mediate protein–protein interactions. Except for the *D. melanogaster* protein CG10631 (accession number AAF53840), in which 27 THAP domains occur consecutively, nearly all known THAP domains are located in the N-terminus of the protein, covering about 90 amino acid residues. In figure 4*A*, the THAP-domain consensus sequence is compared with the corresponding regions of the *P*-homologous proteins of the fruit fly, human, and zebrafish. The conserved sequence motif Cys-$Xaa_{2–4}$-Cys-$Xaa_{35–50}$-Cys-Xaa_2-His, together with the three strictly conserved residues P, W, and F (consensus positions 23, 32, and 55 in figure 4*A*) (Roussigne et al. 2003*b*), can be found in all three proteins. The C-terminal AVPTIF box is highest conserved in the putative Pdre2 protein, whereas several amino acid substitutions occur in the *Drosophila P*-element transposase and in the Phsa protein (fig. 4*A*).

The alignment of the *P*-homologous proteins from human, chicken, and zebrafish as presented in figure 4*B* shows the absence of the THAP domain in the putative Pgga protein. As the presumed start codon of *Pgga* corresponds to the previously postulated start codon of the first *Phsa*-cDNA entry (accession number AK026973), it can be speculated that the respective cDNAs are alternative splicing products and that the THAP domain–containing *Pgga* transcript could not be detected so far. In the case that the THAP domain is really deleted in the Pgga protein, the DNA-binding domain would have been lost, indicating an altered function of the *Pgga* gene product within chicken.

Another known *P*-transposase motif, the leucine zipper, is supposed to play a transposition inhibitory function in *Drosophila* because of dimerization between the transposase and its repressor (Lee, Mul, and Rio 1996). This motif can be found in all *Drosophila P* element–related proteins and comprises the region from aa 101 to aa 122 of the *D. melanogaster* transposase (accession number A24786 [Rio 1990]). This leucine zipper is lacking not only in the domesticated *P*-homologous proteins of human and chicken but also in the presumable transposable zebrafish elements. On the assumption that leucine zipper domains are mediating the repression of transposition, the lack of this domain indicates that activation and repression within the zebrafish genome might follow another mechanism. DNA methylation and demethylation, as well as RNA interference, possibly mediate transposon silencing in various eukaryotic genomes (Kato et al. 2003; Zhou, Cambareri, and Kinsey 2001; Montgomery 2004; Novina and Sharp 2004) and could be the silencing mechanism of *Pdre* elements in the zebrafish genome, too.

Expression Analysis of *Phsa*

To characterize the tissue-specific gene expression pattern of *Phsa*, cDNA panels that comprise cDNA pools of various tissues, including brain (whole), heart, kidney, liver, lung, pancreas, peripheral blood leukocyte, placenta, and skeletal muscle, were subjected to quantitative real-time PCR. The mean calculated concentration of the *Phsa* transcript in the various tissues ranged from 7.983×10^8, as observed in brain, to 1.542×10^{12}, as found in the peripheral blood leukocyte sample (fig. 5*A*). For the HK gene, *G3PDH* the highest level of gene activity was detected in skeletal muscle (3.635×10^{15}), whereas it was the lowest in placenta (1.8×10^{11} [fig. 5*A*]). The significance of the high level of *G3PDH* expression in skeletal muscle was confirmed by 1-way ANOVA (2-tailed, $P = 0.0012$ to 0.0013; t statistic -3.27 to -3.22 [data not shown]) as might be expected

←

FIG. 4.—The THAP domain is an N-terminal protein motif in *P* element–homologous proteins. (*A*) Multiple alignment of the THAP consensus sequence with the THAP domains of the canonical *D. melanogaster P* element transposase (pp25.1; accession number A24786), the human Phsa protein (Phsa; accession number NP_078948), and the putative Pdre2 protein in zebrafish (Pdre2; accession number BX511023). Color code for amino acids between at least two sequences: red = identical amino acids; green = conservative replacements (250 PAMs > 0); and blue = conservative replacements (250 PAMs = 0). Asterisks indicates highly conserved amino acids. (*B*) Multiple alignment of the human (Phsa), the chicken (Pgga), and the zebrafish (Pdre2) *P*-homologous proteins. Boxed amino acids indicate the N-terminal THAP domain. The exact probe region of *Phsa*700-G and *Phsa*700-C used for hybridization experiments is highlighted with the gray box. Color code for amino acids between at least two sequences: red = identical amino acids; green = conservative replacements (250 PAMs > 0 or 250 PAMs = 0).

Table 3
Analysis of Variance of *Phsa* Gene Expression

	Brain	Heart	Kidney	Liver	Lung	Pancreas	Blood	Placenta
Brain	—							
Heart	6.22E-04	—						
Kidney	5.88E–04	−3.41E–05	—					
Liver	−4.40E–03	−5.02E–03	−4.99E–03	—				
Lung	−1.45E–04	−7.67E–04	−7.33E–04	4.25E–03	—			
Pancreas	−9.06E–03	−9.68E–03	−9.65E–03	−4.67E–03	−8.92E–03	—		
Blood	−3.76E–01	−3.76E–01	−3.76E–01	−3.71E–01	−3.76E–01	−3.67E–01	—	
Placenta	−2.34E–02	−2.40E–02	−2.40E–02	−1.90E–02	−2.33E–02	−1.43E–02	3.52E–01	—
Muscle	6.34E–04	1.24E–05	4.65E–05	5.03E–03	7.79E–04	9.70E–03	3.76E–01	2.40E–02

NOTE.—Variance of relative abundance of *Phsa* mRNA in nine human tissue samples was evaluated by 1-way ANOVA. Values represent significant differences. Blood = peripheral blood leukocyte. Muscle = skeletal muscle.

for an enzyme like G3PDH, possible because of the high ATP production in this tissue.

The 1-way ANOVA revealed that the *Phsa* expression in peripheral blood leukocyte was significantly higher than in the other tissues examined (2-tailed, $P = 0.0137$ to 0.0146; t statistic −5.23 to −5.15 [data not shown]). Interestingly, the relative abundance of *Phsa* mRNA differed significantly in all tissue samples, representing peripheral blood leukocyte and skeletal muscle as the landmarks at both ends of the scale (table 3 and fig. 5*B*). However, additional sequence entries in the EST database that are derived from other human tissues (e.g., breast, human testis, colon, cartilage tissue, germinal center B cell from lymphnodes, and embryonic stem cells) strongly indicate that *Phsa* might code for a functional protein that plays a not yet understood but essential role in a specific metabolic pathway.

Molecular Phylogeny of *P*-Homologous Vertebrate Sequences

By means of PCR amplification or Blast searches, we were able to characterize *P*-homologous DNA sequences of almost all analyzed vertebrates (table 2). By using the primer pair Pver1+ and Pver1−, we detected no *P*-homologous DNA sequences in the five rodent species examined: house mouse (*Mus musculus*), black rat (*Rattus rattus*), brown rat (*Rattus norvegicus*), root vole (*Microtus oeconomus*), and the European beaver (*Castor fiber*). Based on paleontological records, a divergence time of about 310 Myr for the separation of birds and mammalian lineages can be assumed, whereas the radiation of the recent mammalian orders dates back approximately 100 Myr (Nelson 1996; Nei and Glazko 2002). Hence, the rodent sequence should have diverged to a lesser extent from other mammals than did the latter from chicken. However, studies focusing on the molecular phylogeny of vertebrates, emphasizing the mammals, argued for an accelerated evolution of rodent genes compared with primate genes (Nei and Glazko 2002; Cotton and Page 2002). To account for the possibly more divergent *P*-homologous sequences in rodents, we used further primer combinations as well as touchdown PCR. Because of the fact that we were able to isolate and detect the highly divergent *P* element homolog from the chicken genome and in zebrafish, respectively, but not from any rodents studied so far, we supposed that the functional *P*-homologous sequence has been lost in the rodent lineage. This interpretation was confirmed by using the UCSC genome browser (mouse assembly, May 2004). On the mouse chromosome 5, positive Blast hits comprised the region from nt 22932957 to nt 22933964 (accession number NT_039308). This sequence corresponds partially to the 3′untranslated region of the *Phsa* gene, suggesting that it is a *P*-homologous rudiment in mouse. Using the cDNA (accession number BX640657), which was used to confirm the orthologous positions of *Phsa* and *Pgga*, as query sequence Blast search revealed that the *P*-homologous rudimentary sequence is located at orthologous position in the mouse genome, too (accession number NT_039308: nt 22939616 to nt 22996577). The same sequence arrangement can be found in rat, where the *P*-homologous rudiment (accession number AABR03089973: nt 27116 to 26115) and the neighboring gene (accession number AABR03089973: nt 1 to nt 17906) are located on chromosome 14. In conclusion, our results suggest that the failure to amplify *P*-homologous sequences in rodents is more likely caused by gene loss than by sequence divergence in this group.

Based on the obtained DNA sequences and their deduced amino acid alignment, we reconstructed the phylogenetic history of the analyzed vertebrate *P*-homologous elements and compared the results with those of the phylogenetic relationships among vertebrates inferred from mt cyt *b* sequences (table 4 and fig. 6*A* and *B*). The reconstructed topologies for the deduced *P*-homologous and the cyt b amino acid data sets are shown in figure 6. The reconstructed phylogenies of both approaches revealed congruence for each data set, although only the trees of the ML analyses of the amino acid data set of each gene are presented (fig. 6*A* and *B*). Comparing the ML tree derived from *P* homologs with that from cyt b, the topologies were not congruent. Moreover, the tree search for the cyt b protein data set revealed a better resolution of the computed tree (i.e., lower number of unresolved quartets [table 4]). Both gene trees clearly confirmed the monophyly of the primates and supported the paraphyletic situation of the insectivores, as well as the aberrant position of the ring-tailed lemur, *Lemur catta*. In the *P*-homologous ML tree (fig. 6*A*), the carnivores formed a monophyletic group, which could not be confirmed by the cyt b ML analysis (fig. 6*B*). Interestingly, the cyt b ML tree displayed a paraphyletic positioning of the rodents (fig. 6*B*). With respect to these results, we can conclude that our *P*-homologous sequence data

Table 4
Estimated Parameters of the Phylogenetic Analyses of Vertebrates

	Maximum-Likelihood Analysis					Neighbor-Joining Analysis			
Data Set	Model	ti/tv[a]	α[a]	Unresolved Quartets[a]	Score (logL)[a]	Model	ti/tv[b]	α[b]	I[b]
Phsa DNA	HKY85	3.25	0.69	591 (9.9%)	−5433.12	F84	3.03	0.68	0.00
Phsa protein	Dayhoff78	—	—	643 (10.7%)	−2857.06	JTT92	—	—	—
cyt b DNA	TrN93	2.49	0.27	1350 (10.7%)	−14894.20	F84	—	0.52	0.33
cyt b protein	mtREV24	—	—	212 (1.7%)	−5647.70	JTT92	—	—	—

NOTE.—ti/tv = transition/transversion ratio; α = gamma distribution shape parameter; I = proportion of invariable sites.
[a] Parameters were estimated from the respective data sets.
[b] Parameters were estimated from the respective data sets by using the program Modeltest (Posada and Crandall 1998).

reflect to a certain extent the phylogenetic relationships among vertebrates (i.e., ML analyses of entire mitochondrial genomes [Arnason et al. 2002; Nikaido et al. 2001]). In conclusion, the phylogeny of the *P*-homologous sequences from the vertebrates studied so far is more or less in accordance with the species phylogeny, indicating their vertical transmission.

Molecular Domestication History of *P*-Homologous Sequences

Two hypotheses have been discussed to explain active *P* element transposons in some dipteran species and stationary *P* element transposon–derived sequences in others. The "domestication hypothesis" considers those stationary sequences as degenerated transposons that have lost the structural features necessary for transposition like terminal inverted repeats and, thus, are not able to move any more. In some cases, these now immobile transposon derivates have acquired a novel function and represent a stable functional component within their host genome (Pinsker et al. 2001). In contrast, the "source gene hypothesis" argues for an evolutionary scenario in which active *P* element transposons are direct descendants of an ancient source gene that gave rise to a transposon by acquisition of terminal inverted repeats.

In zebrafish, *P*-homologous sequences are spread throughout the genome, indicating their transposable activity and contributing to two different scenarios that can be observed within the zebrafish genome: the terminally truncated *Pdre* sequences represent immobile components of the genome, whereas *Pdre2* and the analyzed internally deleted copies show structural features typical for DNA transposons. Quite contrary features can be described for *Phsa* in human and *Pgga* in chicken, where they represent stationary single-copy genes. *Phsa*, *Pgga*, and the *P*-homologous mouse and rat rudiments are located at orthologous positions, thus, favoring the "source gene hypothesis." Although the immobile terminally truncated *Pdre* elements are located at paralogous positions compared with *Pgga* and the mammalian *P*-homologous sequences, the data obtained from zebrafish do not support the "source gene hypothesis." They indicate that *P*-homologous sequences as immobile and subsequently stable components of the genome were generated by "molecular domestication." As a consequence of this explanation, the domestication event of the *Phsa*/*Pgga* sequence has occurred by

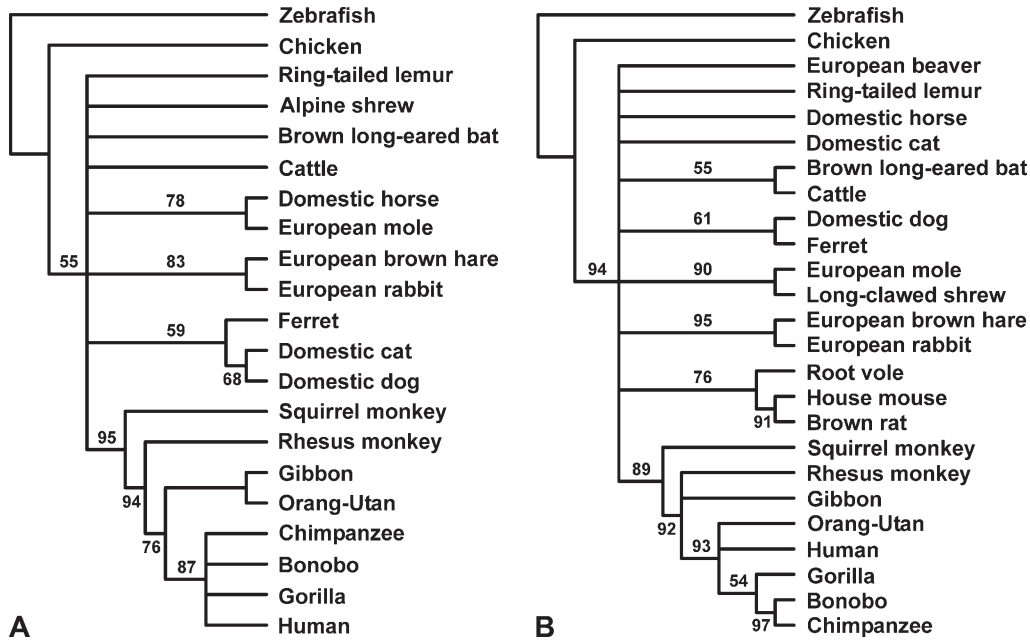


FIG. 6.—The reconstructed maximum-likelihood trees of studied vertebrates based on the homologous amino acid sequences of the nuclear *Phsa* gene (*A*) and the mitochondrial *cyt b* gene (*B*). Numbers at the nodes refer to percentage reliability values of quartet puzzling, corresponding to the number of times the group appears after 10,000 puzzling steps. For more details see the multiple sequence alignments in Supplementary Material online.

immobilization of an active transposon within a common ancestor before the separation of mammalians and birds about 310 MYA (Nelson 1996).

P-homologous sequences are not the only examples of domesticated transposons. A considerable number of domesticated transposable-element copies now contribute to transcriptionally regulatory elements or to protein-coding regions of cellular genes (Smit 1999, International Human Genome Sequencing Consortium 2001, Jordan et al. 2003). The high number of known transposon-derived domesticated genes provides further evidence that the stable vertebrate *P*-homologous sequences were recruited as novel genes from their respective host genomes by "molecular domestication" of a former active transposon copy.

Supplementary Material

The sequences reported in this paper have been submitted to DDBJ/EMBL/GenBank database and have been assigned accession numbers AJ717666 to AJ717685. The accession numbers of the vertebrate mtDNA *cyt b* gene sequences plus the input files for the multiple sequence alignments (*P-homologs-infiles.txt*), the P-homologs-DNA, P-homologs-protein, cyt b DNA, and cyt b protein ClustalX 1.81 multiple sequence alignments are available at the MBE Web site. The DNA alignment file for the phylogenetic analysis of the *P*-homologous vertebrate sequences and the alignment file of the mitochondrial cytochrome b protein sequences are deposited in EMBL-Align database with accession numbers ALIGN_000704 and ALIGN_000703, respectively.

Acknowledgment

The German Primate Center (Göttingen, Germany) provided valuable DNA of the primates analyzed in this study. We thank Alma Sendic (Center of Physiology and Pathophysiology at the Medical University Vienna, Austria) for providing chemicals, the LightCycler instrument, and her excellent knowledge of quantitative PCR analysis. We are indebted to Katrina Vanura (Division of Hematology at the Medical University Vienna, Austria) for various inspirations and discussions to improve our experimental protocols. We are also grateful to Ingrid Gerstl and Traude Kehrer for their outstanding laboratory assistance. The helpful comments of two anonymous reviewers on previous versions of this paper are greatly acknowledged. The study was funded by the Austrian Science Foundation (FWF project no. P15785).

Literature Cited

Adachi, J., and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

Analyse-it for Microsoft Excel. 2003. Version 1.71. Software Ltd., Leeds, UK.

Arnason, U., J. A. Adegoke, K. Bodin, E. W. Born, Y. B. Esa, A. Gullberg, M. Nilsson, R. V. Short, X. Xu, and A. Janke. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. Proc. Natl. Acad. Sci. USA **99**:8151–8156.

Ausubel, F. A., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl (eds.) 1998. Current protocols in molecular biology. John Wiley and Sons, New York.

Cotton, J. A., and R. D. M. Page. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Lond. B Biol. Sci. **269**:1555–1561.

Daniels S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics **124**:339–355.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed. Atlas of protein sequence structure, Vol. 5. National Biomedical Research Foundation, Washington, DC.

Farrer, T., A. B. Roller, W. J. Kent, and A. M. Zahler. 2002. Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing. Nucleic Acids Res. **30**:3360–3367.

Felsenstein J. 2004. PHYLIP (phylogeny inference package). Version 3.6b. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Felsenstein J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

Gale J. R., M. Blakely, C. M. Hopkins, D. A. Melville, M. W. Wambach, M. Romano, P. R. Katze, and G. Michael. 1998. Regulation of interferon-induced protein kinase PKR: modulation of $P58^{IPK}$ inhibitory function by a novel protein, $P52^{rIBK.}$ Mol. Cell. Biol. **18**:859–871.

Genetic Analysis Manual: Global Edition IR^2 System. 1999. Publication Number 9910-117, LI-COR, Inc. Lincoln, Nebrasha, USA.

Hagemann, S., and W. Pinsker. 2001. *Drosophila P* transposons in the human genome? Mol. Biol. Evol. **18**:1979–1982.

Hagemann, S., E. Haring, and W. Pinsker. 1996. Repeated horizontal transfer of *P* transposons between *Scaptomyza pallida* and *Drosophila bifasciata*. Genetica **98**:43–51.

Haring, E., S. Hagemann, and W. Pinsker. 2000. Ancient and recent horizontal invasions of drosophilids by *P* elements. J. Mol. Evol. **51**:577–586.

Hasegawa, M., H. Kishino, and K. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

International Human Genome Sequencing Consortium.2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Jordan, I. K., I. B. Rogozin, G. V. Glazko, and E. V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. **19**:68–72.

Kato, M., A. Miura, J. Bender, S. E. Jacobsen, and T. Kakutani. 2003. Role of CG and non-CG methylation in immobilization of transposons in *Arabidopsis*. Curr. Biol. **13**:421–426.

Kidwell, M. G., J. F. Kidwell, and J. A. Sved. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics **86**:813–833.

Koenig, M., M. Reichel, R. Marschalek, O. A. Haas, and S. Strehl. 2002. A highly specific and sensitive fluorescence in situ hybridization assay for detection of t(4;11)(q21;q23) and concurrent submicroscopic deletions in acute leukaemias. Brit. J. Haematol. **116**:758–764.

Lee, S. H., J. B. Clark, and M. G. Kidwell. 1999. A *P* element-homologous sequence in the house fly, *Musca domestica*. Insect Mol. Biol. **8**:491–500.

Marchler-Bauer, A., J. B. Anderson, C. DeWeese-Scott, et al. (24 co-authors).2003. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. **31**:383–387.

Miller, W. J., S. Hagemann, E. Reiter, and W. Pinsker. 1992. *P*-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. Proc. Natl. Acad. Sci. USA **89**:4018–4022.

Miller, W. J., J. F. McDonald, D. Nouaud, and D. Anxolabéhère. 1999. Molecular domestication—more than a sporadic episode in evolution? Genetica **107**:197–207.

Misra, S., and D. C. Rio. 1990. Cytotype control of *P* element transposition: the 66 kD protein is a repressor of transposase activity. Cell **62**:269–284.

Montgomery, M. K. 2004. RNA interference: historical overview and significance. Methods Mol. Biol. **265**:3–21.

Nei, M., and G. V. Glazko. 2002. Estimation of divergence times for a few mammalian and several Primate species. J. Hered. **93**:157–164.

Nekrutenko A., and W.-H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. **17**:619–621.

Nelson D. R. 1996. Molecular evolution A lecture series by David R. Nelson, placed on the Internet on 21 November, 1996. http://www.icp.ucl.ac.be/~opperd/private/m_e_index.html.

Nikaido, M., K. Kawai, Y. Cao, M. Harada, S. Tomita, N. Okada, and M. Hasegawa. 2001. Maximum likelihood analysis of the complete mitochondrial genomes of the Eutherians and a reevaluation of the phylogeny of bats and insectivores. J. Mol. Evol. **53**:508–516.

Nouaud, D., and D. Anxolabéhère. 1997. *P* element domestication: a stationary truncated *P* element may encode a 66-kDa repressor-like protein in the *Drosophila montium* species subgroup. Mol. Biol. Evol. **14**:1132–1144.

Novina, C. D. and P. A. Sharp. 2004. The RNAi revolution. Nature **430**:161–164.

O'Hare, K., and G. M. Rubin. 1983. Structures of *P* transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell **34**:25–35.

Oliveira de Carvalho, M., J. C. Silva, and E. L. Loreto.2004. Analyses of *P*-like transposable element sequences from the genome of *Anopheles gambiae*. Insect Mol. Biol. **13**:55–63.

Perkins, H. D., and A. J. Howells. 1992. Genomic sequences with homology to the *P* element of *Drosophila melanogaster* occur in the blowfly *Lucilia cuprina*. Proc. Natl. Acad. Sci. USA **89**:10753–10757.

Pinsker, W., E. Haring, S. Hagemann, and W. J. Miller. 2001. The evolutionary life history of *P* transposons: from horizontal invaders to domesticated neogenes. Chromosoma **110**:148–158.

Posada, D. and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **17**:17–818.

Rio, D. C. 1990. Molecular mechanism regulating *Drosophila* P element transposition. Annu. Rev. Genet. **24**:543–578.

Roussigne M., C. Cayrol, T. Clouaire, F. Amalric, and J. P. Girard. 2003*a*. THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. Oncogene **22**:2432–2442.

Roussigne M., S. Kossida, A. C. Lavigne, T. Clouaire, V. Ecochard, A. Glories, F. Amalric, and J. P. Girard. 2003*b*. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of *P* element transposase. Trends Biochem. Sci. **28**:66–69.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:1406–1425.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning—a laboratory manual. 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sarkar, A., R. Rengupta, J. Krzywinski, X. Wang, C. Roth, and F. H. Collins. 2003. *P* elements are found in the genomes of nematoceran insects of the genus *Anopheles*. Insect Biochem. Mol. Biol. **33**:381–387.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502–504.

Silva, J. C., and M. G. Kidwell. 2000. Horizontal transfer and selection in the evolution of *P* elements. Mol. Biol. Evol. **12**:391–404.

Smit A. F. A. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9**:657–663.

Sulston, J. E., and R. Waterston.1998. Toward a complete human genome sequence. Genome Res. **8**:1097–1108.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**:4876–4882.

Zhou, Y., E. B. Cambareri, and J. A. Kinsey. 2001. DNA methylation inhibits expression and transposition of the Neurospora Tad retrotransposon. Mol. Genet. Genomics **265**:748–754.

Lauren McIntyre, Associate Editor

Accepted December 14, 2004