

Analysis of Nuclear Receptor Pseudogenes in Vertebrates: How the Silent Tell Their Stories

Zhengdong D. Zhang,* Philip Cayting,* George Weinstock,† and Mark Gerstein*‡§

*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut; †Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas; ‡Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut; and §Department of Computer Science, Yale University, New Haven, Connecticut

Transcription factor pseudogenes have not been systematically studied before. Nuclear receptors (NRs) constitute one of the largest groups of transcription factors in animals (e.g., 48 NRs in human). The availability of whole-genome sequences enables a global inventory of the NR pseudogenes in a number of vertebrate model organisms. Here we identify the NR pseudogenes in 8 vertebrate organisms and make our results available online at <http://www.pseudogene.org/nr>. The assignments reveal that NR pseudogenes as a group have characteristics related to generation and distribution contrary to expectations derived from previous large-scale pseudogene studies. In particular, 1) despite its large size, the NR gene family has only a very small number of pseudogenes in each of the vertebrate genomes examined; 2) despite the low transcription levels of NR genes, except for one, all other NR pseudogenes identified in this study are retropseudogenes; and 3) no duplicated NR pseudogenes are found, contrary to the fact that the NR gene family was expanded through several waves of gene duplication events. Our analyses further reveal a number of interesting aspects of NR pseudogenes. Specifically, through careful sequence analysis, we identify remnant introns in 2 mouse retropseudogenes, ψ Rev-erb β and ψ LRH1. Generated from partially processed pre-mRNAs, they appear to be rare examples of highly unusual “semiprocessed” pseudogenes. Second, by comparing the genomic sequences, we uncover a pseudogene that is unique to the human lineage relative to chimpanzee. Generated by a recent duplication of a segment in the human genome, this pseudogene is a “duplicated–processed” pseudogene, belonging to a new pseudogene species. Finally, *FXR β* was nonfunctionalized in the human lineage and thus appears to be an example of a rare unitary pseudogene. By comparing orthologous sequences, we dated the *FXR–FXR β* duplication and the nonfunctionalization of *FXR β* in primates.

Background

Nuclear receptors (NRs) regulate nuclear gene expression in response to various extracellular and intracellular signals and play a prominent role in a group of diverse and critical biological processes such as reproduction, differentiation, development, metabolism, metamorphosis, and homeostasis. Activated by binding of small hydrophobic molecules, they provide a direct link between ligands that signal different stages of those processes and cells' transcriptional responses. All NRs share a similar domain arrangement and, with a few exceptions, contain both the DNA-binding domain (DBD) and the ligand-binding domain (LBD), the 2 most conserved signature domains of this protein family. NRs have been specifically surveyed and studied in several species whose genomes have been fully sequenced, which include *Ciona intestinalis* (Dehal et al. 2002), *Caenorhabditis elegans* (Sluder et al. 1999), *Drosophila melanogaster* (Adams et al. 2000), human (Robinson-Rechavi et al. 2001; Zhang et al. 2004), mouse (Zhang et al. 2004), and rat (Zhang et al. 2004).

Pseudogenes (ψ) are nongenic DNA segments that exhibit a high degree of sequence similarity to functional genes but contain disruptive defects, including, not exhaustively, premature stop codons, splice site mutations, and frameshift mutations, which prevent them from being expressed properly. Disruption in the promoter regions of gene can also result in its pseudogenization. Based on whether they have gone through RNA processing,

pseudogenes can be classified into 2 categories: processed and unprocessed pseudogenes. Processed pseudogenes are generated by the integration of the reverse transcription products of processed mRNA transcripts into the genome. Unprocessed pseudogene has not gone through RNA processing and thus has retained the original exon–intron structure of the functional gene.

Previous studies have identified 3 NR pseudogenes in human: ψ ERR α (Sladek et al. 1997), ψ HNF4 γ (Tchenio et al. 1993), and ψ FXR β (Maglich et al. 2001; Otte et al. 2003) (see table 1 for symbols and full names of NRs included in this study). Recently, several other NR pseudogenes were also identified in mice and rats (Zhang et al. 2004). However, the availability of 8 vertebrate genome sequences (Waterston et al. 2002; Gibbs et al. 2004; International Chicken Genome Sequencing Consortium 2004; International Human Genome Sequencing Consortium 2004; The Chimpanzee Sequencing and Analysis Consortium 2005; Lindblad-Toh et al. 2005) makes it possible to conduct a detailed study of the NR pseudogenes in both human and vertebrate model systems. Here we present a comprehensive survey of NR pseudogenes in these 8 vertebrate genomes and report their locations, sequences, and defects. Recently, pseudogenes in the entire human genome have been identified either in gene family–specific studies (Glusman et al. 2001; Zhang et al. 2002) or in comprehensive surveys (Ohshima et al. 2003; Torrents et al. 2003; Zhang et al. 2003). Based on the mechanisms for pseudogene generation and the observations reported in those large-scale studies, we expected that NR pseudogenes would be mostly duplicated pseudogenes (like olfactory receptor pseudogenes) and few processed ones as NR genes were created by multiple gene duplication events and most NR genes have low expression levels. Our survey results here, however, are in striking opposition to these initial

Key words: nuclear receptor, pseudogene, nonfunctionalization, protein evolution.

E-mail: mark.gerstein@yale.edu.

Mol. Biol. Evol. 25(1):131–143. 2008

doi:10.1093/molbev/msm251

Advance Access publication December 7, 2007

Table 1
Symbols of NR Used in the Text

Symbol	Official Name	Full Name
FXR β	NR1H5	Farnesoid X receptor, beta
HNF4 γ	NR2A2	Hepatocyte nuclear factor 4, gamma
ERR α	NR3B1	Estrogen-related receptor, alpha
Rev-erb β	NR1D2	Thyroid hormone receptor, alpha-like
PNR	NR2E3	Photoreceptor-specific NR
ERR β	NR3B2	Estrogen-related receptor, beta
LRH1	NR5A2	Liver receptor homolog 1

expectations. The analysis of these pseudogenes affords unique insights into the evolution and dynamics of this gene family and the mammalian genomes at large.

Methods

The human, mouse, and rat genomic sequences used in this study were human genome build of May 2004, mouse genome build of May 2004, and rat genome build of June 2003. Each of these 3 genomes was partitioned into 750-kb segments with 2-kb overlaps to take advantage of parallel computing. The DBD and LBD (designated as zf-C4 and hormone_rec in the Pfam database) were searched in the genomic sequences using GENEWISEDB. Predictions with frameshifts and premature stop codons that could not be credibly attributed to the sequencing errors were retained and aligned with 62 representative NR protein sequences to reveal their identities, which were the best BlastP hits. NR protein sequences to which these predictions were identified were then aligned to 10-kb genomic sequence intervals centered on the positions of these predictions using both GENEWISEDB and BLAT. The sequences, defects, and structures of the NR pseudogenes were constructed from GENEWISEDB and BLAT alignments, which verified and complemented each other.

To estimate the date of *FXR*–*FXR β* duplication (T_D), 4 homologous sequences, *FXR*_{mouse}, *FXR β* _{mouse}, *FXR*_{rat}, and *FXR β* _{rat}, were used (Li 1997). Because the synonymous substitutions per synonymous site (K_s) are large and thus cannot be estimated accurately, they are not used to calculate T_D . As the equation shows below, only the nonsynonymous substitutions per nonsynonymous site (K_a) are used. T_D is estimated by

$$T_D = 2 \cdot T_S \cdot \frac{K_a \text{ } FXR, FXR\beta}{K_a \text{ } FXR + K_a \text{ } FXR\beta},$$

where T_S is the divergence time between mouse and rat, for which 41 Myr were used in the calculation (Hedges 2002), is the average value of 4 numbers of nucleotide substitutions per site estimated from 4 pairwise comparisons: *FXR*_{mouse}–*FXR β* _{mouse}, *FXR*_{mouse}–*FXR β* _{rat}, *FXR*_{rat}–*FXR β* _{mouse}, and *FXR*_{rat}–*FXR β* _{rat}, and $K_a \text{ } FXR$ and $K_a \text{ } FXR\beta$ are the numbers of the synonymous substitutions per synonymous site in *FXR* and *FXR β* , respectively (supplementary table 1, Supplementary Material online).

To estimate the nonfunctionalization time (T_N) of ψ *FXR β* in the primate lineage, we used the method devised by Chou et al. See Chou et al. (2002) for a detailed descrip-

tion of the method. Briefly, it assumes that nonsynonymous mutations are selected against until the gene is inactivated; thereafter, mutations at both synonymous and nonsynonymous sites accumulate at the neutral mutation rate. Quantification of lineage-specific mutation rates at synonymous and nonsynonymous sites remote from the inactivating deletion provides the information necessary for the calculation. Four *FXR β* sequences, from human, mouse, rat, and chicken, were used for the calculation (supplementary table 2, Supplementary Material online). We used the method proposed by Li et al. (1981) to estimate the non-functionalization time of all retropseudogenes identified in this study. Because they are “dead on arrival,” we assumed that $T_N = T_D$.

Multiple *FXR* and *FXR β* peptide sequences together with the human *LXR α* peptide sequences were aligned using MUSCLE (Edgar 2004). The phylogeny of *FXR* and *FXR β* was constructed from this sequence alignment using an implementation of the Neighbor-Joining algorithm in the PAUP*4.0 software package with a bootstrap of 1,000 replicates. The tree was rooted by *LXR α* .

Results

NR Pseudogenes in Vertebrate Model Organisms

By using manual annotation and a pseudogene identification pipeline, we assigned NR pseudogenes in human, chimpanzee, mouse, rat, dog, chicken, tetraodon, and zebrafish—8 vertebrate model organisms whose genomes have been sequenced. Our identification results are available at <http://pseudogene.org/nr>. We focused our analyses on NR pseudogenes in human, chimpanzee, mouse, and rat due to the incomplete genome annotation for the other vertebrate genomes, which prevents complete assignments and confident interpretation of pseudogenes identified in those genomes. However, as the annotation improves, we will update our NR pseudogene assignments and post the results online.

Overall, there are only a very small number of NR pseudogenes in each of the vertebrate genomes examined. Within the human, chimpanzee, mouse, and rat genomes, 4, 3, 5, and 3 NR pseudogenes were identified, respectively (table 2). The existence of the 3 previously reported pseudogenes in the human genome— ψ *ERR α* (Sladek et al. 1997), ψ *HNF4 γ* (Tchenio et al. 1993), and ψ *FXR β* (Maglich et al. 2001; Otte et al. 2003)—was confirmed by our analysis. Except for 1 human NR pseudogene, ψ *FXR β* , which is unprocessed, all other NR pseudogenes identified are retropseudogenes. No duplicated NR pseudogenes were identified, a finding quite contrary to our expectation as described above and in the discussion—that is, because NR genes encode transcription factors and generally have low and restricted transcription profiles, we expected most of the NR pseudogenes to be created by duplication.

Two ψ *ERR α* Are in the Human Genome

Sladek et al. (1997) reported the isolation of a processed *ERR α* pseudogene mapped to human chromosome

Table 2
Human and Rodent NR Pseudogenes

Genome	Pseudogene	Accession ^a	Location ^b		Type	Truncation ^c	
			Chromosome	Band		5'	3'
Human	ψ FXR β	15259	1p13.1+	115181466	Unitary	No	No
	ψ HNF4 γ	128390	13q21.1–	55471366	Processed	Yes	Yes
	ψ ERR α	5316	13q12.11–	19032156	Processed	No	No
	ψ ERR α	24162	13q12.11+	20732460	Processed	No	No
Chimpanzee	ψ FXR β	8400	1–	122802667	Unitary	No	No
	ψ HNF4 γ	8401	13–	55892954	Processed	Yes	Yes
	ψ ERR α	8402	13–	19079069	Processed	No	No
	ψ Rev-erb β	19393	19qC3+	40244011	Semiprocessed	No	No
Mouse	ψ PNR	6324	15qB3.1+	35678192	Processed	Yes	No
	ψ ERR β	10804	XqA5+	57351250	Processed	No	No
	ψ LRH1	8260	3qH2+	144716412	Semiprocessed	Yes	No
	ψ LRH1	17110	6qF1–	118583245	Processed	Yes	No
Rat	ψ ERR β	8720	Xq36+	146717523	Processed	No	No
	ψ LRH1	1916	11q21+	48578386	Processed	Yes	No
	ψ LRH1	17561	Xq14–	30976310	Processed	Yes	No

^a The pseudogene accession numbers as in the Yale Pseudogene Database. Prefix the number with “urn:lsid:pseudogene.org:9606.Pseudogene”: to get the whole accession key. Visit <http://www.pseudogene.org> for details.

^b The genomic location indicates the chromosome band (only the chromosome number and strand for the chimpanzee genome as other band information is currently not available), the strand (“+” being forward and “–” reverse), and the start coordinate of the pseudogene sequence in the genome. The reference genomes are human of March 2006 (Hsap NCBI Build 36.1, hg18), chimpanzee of March 2006 (panTro2), mouse of February 2006 (Mmus NCBI Build 36, mm8), and rat of November 2004 (Rnor3.4), respectively.

^c Truncation is relative to the coding sequences; 5' and 3' refer to the ends of the coding sequence of the functional parent gene.

13q12.1. In our study, however, 2 processed ψ ERR α s (ψ ERR α – and ψ ERR α –), immediately next to each other on opposite DNA strands, were identified in the same chromosome band (13q12.11). The genomic sequence interval between these 2 ψ ERR α , approximately 1.7 Mb, is well below the maximum resolution of conventional fluorescence in situ hybridization used by Sladek et al. on metaphase chromosomes and thus precluded the identification of both of pseudogenes in their study.

These 2 human ψ ERR α sequences are very similar (but not identical, which rules out the possibility of a sequence assembly error): their Hamming distance, D_H , which measures the proportion of site differences between 2 sequences, is only 3.65% and the number of nucleotide substitution per site between them, K , is 0.038 ± 0.006 . The ψ ERR α on the forward strand contains 5 frameshifts, the ψ ERR α on the reverse strand has 4, and both have a premature stop codon at different positions. Of these defects in their sequences, 3 frameshifts are identical. Except for several internal deletions, both ψ ERR α are of full length and highly similar, albeit defunct, copies of the transcript of the functional gene, which suggests a young age (~ 38 MYA) for both of them.

As expected, we identified a set of NR pseudogenes in chimpanzee similar to those in human. However, the chimpanzee ortholog of the human ψ ERR α – is absent. This absence indicates that ψ ERR α – was created first, at least before the divergence of human and chimpanzee, and at the same time the high sequence similarity and the shared defects between human ψ ERR α – and ψ ERR α – suggest that the former was created by the duplication of the latter in the human lineage after its divergence from chimpanzee. In fact, those 2 pseudogenes reside in 2 expansive (> 14.6 kb) and highly similar (96% identical) sequence segments in the human chromosome 13 that were created by a recent (< 6

MYA), human-specific segmental duplication (Bailey et al. 2002; Cheng et al. 2005). Thus, human ψ ERR α – is a duplication of a processed pseudogene. This “duplicated–processed” pseudogene belongs to a new category of pseudogenes—first noted in a study of the human cytochrome *c* pseudogenes (Zhang and Gerstein 2003)—that are different from either duplicated or processed pseudogenes in terms of their underlying generating processes. The original processed pseudogene and the pseudogene duplicated from it both have little consequence to the fitness of the organism. Nevertheless, they are distinct pseudogene species. The distinction made between them is important for estimating the frequency of retrotransposition of mRNA transcripts. Clearly, such estimation will be inflated if the duplicated–processed pseudogenes are not excluded as they were generated by duplication, not retrotransposition, events.

Human ψ FXR β Is a Unitary Pseudogene with Multiple Nonfunctionalization Mutations

Previous studies (Maglich et al. 2001; Otte et al. 2003) have shown that human FXR β is an unprocessed pseudogene with no functional counterpart (unitary pseudogene) in the human genome. This gene was also nonfunctionalized in other Old World primates studied so far but encodes a functional receptor in other mammals (see Otte et al. [2003] and below). The alignment of the mouse FXR β protein sequence to the 3-frame translation of the human genomic sequence reveals that the “coding sequence” of the original human FXR β gene were interrupted by at least 9 introns, and in the currently defunct gene, there are 10 disruptive defects, which consist of 4 frameshifts, 4 nonsense mutations, and 3 splice site mutations (fig. 1). These defects are equally distributed at the beginning and the end of this pseudogene.

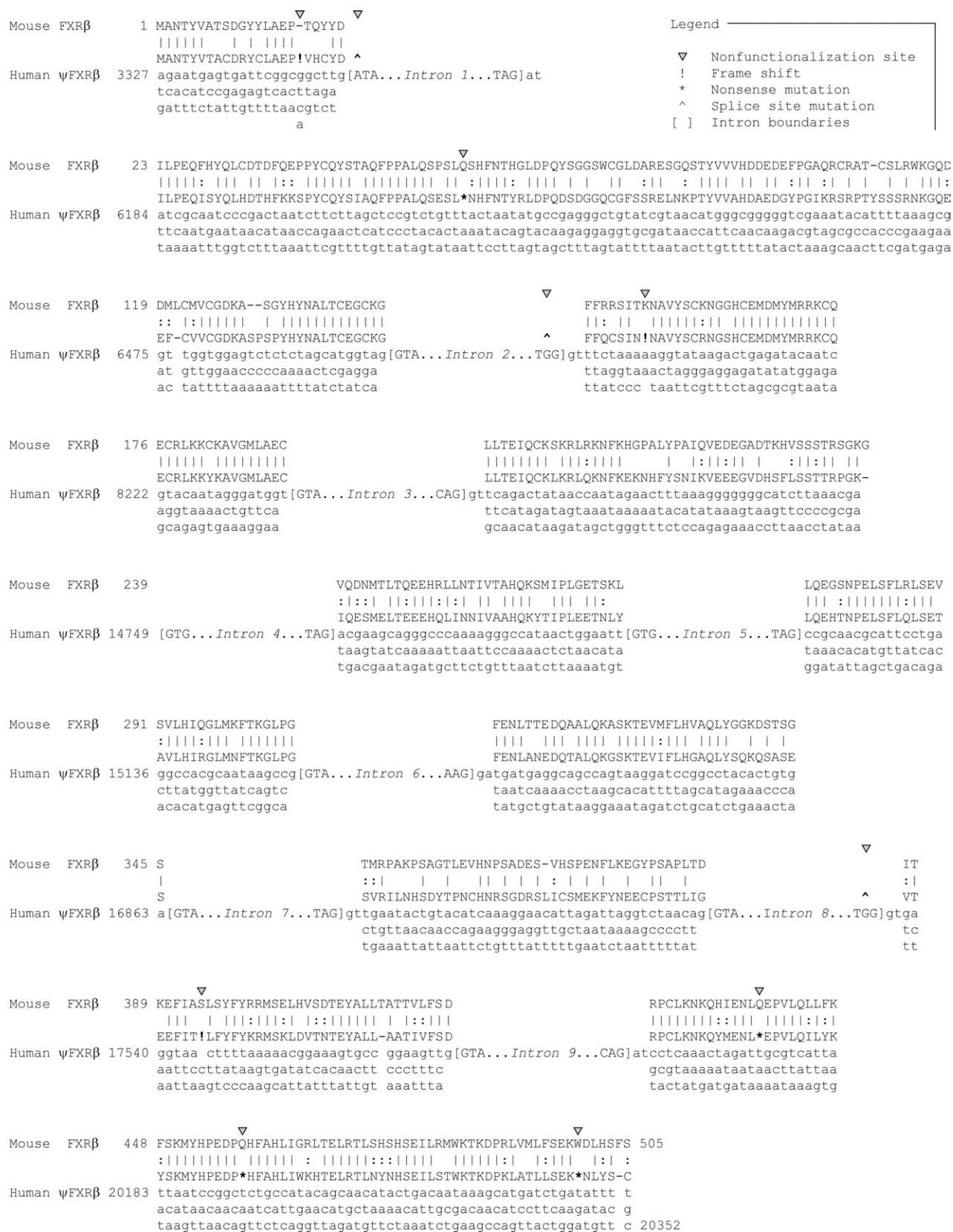


Fig. 1.—The gene structure of human ψ FXR β . The mouse FXR β protein sequence and the translation of the human genomic sequence at the ψ FXR β locus are aligned. The identical and similar character states in the alignment are indicated by vertical lines and colons, respectively. The identified sequence defects in human ψ FXR β locus are denoted in its translation by different symbols according to their types (see the figure key table) and also marked uniformly above the alignment. The human sequence coordinates indicate the distance of the nucleotide from the beginning of the genomic sequence from the sequencing clone RP11-350E19 (GenBank accession: AL358372.11).

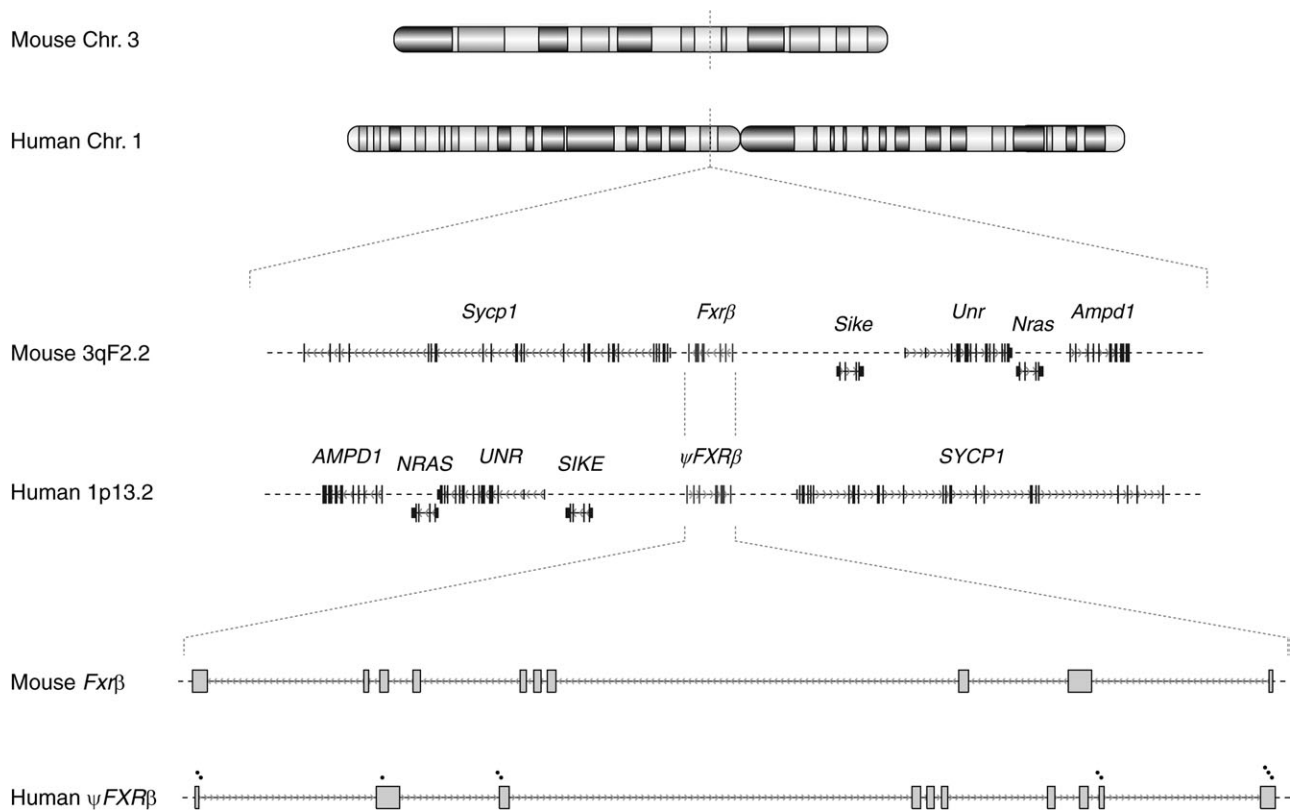


FIG. 2.—The genomic context of human and mouse ψ FXR β loci. The gene structure was constructed from the sequence alignment of mouse FXR β protein sequence to the translated human genomic sequence. The approximate locations of the defects in human ψ FXR β are indicated by black dots above its enlarged gene structure. All exons, introns, and intergenic regions are drawn in proportion.

Human ψ FXR β and its mouse ortholog are located in 2 expansive (>25 Mb) syntenic regions in the 2 genomes (fig. 2). The same set of genes, in an identical order and orientation, in 2 genomic neighborhood make it unlikely that human FXR β was inactivated by a chromosomal translocation or other genomic rearrangement processes. The comparison of the orthologous sequences from human, chimpanzee, and rhesus (fig. 3A) reveals both ancestral and lineage-specific sequence defects, 14 in all, in ψ FXR β from these 3 primates (fig. 3B). The disruptive mutations at the 1st, 2nd, and 14th positions in ψ FXR β are present in all 3 species and, hence, most likely arose in the common ancestor of human, chimpanzee, and rhesus. Because the mutation at the 14th position, a nonsense mutation, is at the very end of the coding sequence and thus had considerably less disrupting power, either of the other 2 common mutations, 1 frameshift mutation and 1 splice site mutation at the start of the reading frame, could be the mutation that pseudogenized FXR β in these primates. The orthologous genomic sequences from other primate species would make it possible to pin down the silencing mutation.

Based on 4 pairwise comparisons among the mouse and rat FXR and FXR β sequences, our study dated the ancient gene duplication event that created this pair of paralogous genes to be ~496 MYA prior to the speciation events (~450 MYA) that ultimately gave rise to fishes and other vertebrates (fig. 4A). This estimation was confirmed by the search result for FXR and FXR β in the genomes of repre-

sentative species that both genes exist in human, chimpanzee, mouse, chicken, frog (*Xenopus tropicalis*), and fish (both zebrafish and puffer fish, supplementary fig. 1, Supplementary Material online). The phylogeny of FXR and FXR β reveals that by the measure of branch length (data not shown), FXR β is evolving at least 5.6 times faster than FXR in mammals, but a similar difference in the evolution speed is not observed in nonmammalian vertebrates (fig. 4B, see supplementary fig. 2, Supplementary Material online for the multiple sequence alignment). Based on human, mouse, rat, and dog FXR β sequences, our calculation indicates that the silencing of FXR β happened ~42 MYA.

Intergenic Sequences Immediately Upstream and Downstream to Human ψ FXR β Are Conserved

Human ψ FXR β is a transcribed pseudogene: Real-time quantitative polymerase chain reaction detected relatively high levels of expression of its mRNA in testis (Maglich et al. 2001; Otte et al. 2003). This strongly suggests that the promoter and possibly other *cis*-acting elements that regulate the transcription of human ψ FXR β have remained largely intact and functional even long after the inactivation of ψ FXR β . Alignment of multiple genomic sequences from 14 vertebrates including human shows strong sequence conservation in the upstream noncoding regions—where regulatory elements may reside—of

A

			(1)	(2)
Human	ψFXRβ	atggcaaatacttatgtcactgcatgtgataggattgtcttgcgaaccag	<u>gtca</u> gtgcattgctatg	[ATA... <i>Intron</i>
Chimp	ψFXRβ		<u>gtca</u>	[A].....
Rhesus	ψFXRβ		<u>gtca</u>	[A].....

n1...TAG]atattttaccagaacaaattagttatcagctgcatgacactcattttaaaaaatcaccttattgccagttattctattgctcagtttctctcc
.....].
.....

(3)
agctttacagctctgaatctttataaaatcatttcaacacttatagattggatccacaggacagtgatggtggacagtggtgatttagttctcgtagaatta
taa.g.
.....

aataaaccacttatgtggttgcctcatgatgctgaagatggataccctggaataaaaagggtccagaccaacctattcttctcgcagaaataagggacaggy
.....

aagaattctgtgtagtttgtggtgataaagcatcaccatcaccatatacattataatgcacttacctgtgaaggttgcaaag[GTA...*Intron2*...TG
.....g.....[.....G
.....

(5)
G]gtttttttcaatgtagcatcaacaaaatgcagttatatagttgcaggaaatggtagtcactgtgaaatggacatgtacatgcgtagaaaaatgtcaagagt
.]...C.....aa.....g.....
.....

(6)
gcagactgaaaaagttataaggcagtaggaatgttggcagaat[GTA...*Intron3*...CAG]gtttgctcacagaaatccaatgtaaattaaagagact
.....g.....[.....]
.....G].....cg.g.....

(7)
tcaaaagaactttaaggagaagaatcatttttactctaacaatcaaagtggaagaggaaggagtagaccacagttttctatcatccaccactagacctgga
.....a.....
.....C...t...tga.....g.....g.....a...ca.....g...t.....

(8)
aaa[GTG...*Intron4*...TAG]attcaggaagcatggaactaactgaagaggaacatcagctcattaataacattgtggctgctcatcaaaaaatata
...[A].....]
...[.....C.....]
.....

ccattccttttagaagaacaaatttgat[GTG...*Intron5*...TAG]ctgcaggaacatacaaatcctgaactgagctttttgcaactctcagagac
.....[.....].....g.....
.....t.[.....g.....t.....t.....

agcagtcctacacatacgtgggctaataatgaattttaccaaggggtcccag[GTA...*Intron6*...AAG]gatttgaaaatttggccaatgaggatcaa
.....[.....]
.....t.....a.....[.....a.g.....

actgcactacagaagggatcaaaaactgaagtgatatttctccatggggcccaactttacagtcagaaacaatcagcctctgaaa[GTA...*Intron7*...
.....[.....]
.....g.....C.....g.....a.....[.....

..TAG]gttctgtgagaatattaaatcattcagattatacaccaaattgtcacaaataggagtggatagaagtctatttgttctatggaaaaatttta
.....]
.....]a.....tg.....g.....C.....C.....g...

(9) (10)
caatgaagaatgtccttctactactctaattg[GTA...*Intron8*...TG]gtgttactgaagaattttatac-acactgttttacttctacaaaaga
.....[.....G.].....ac.....
.....a.....C.....[.....A.]...a.....C.....g.....

atgagcaaaccttgatgtaactaataactgaatatgctctgcttgcagcaacaattgttttttcag[GTA...*Intron9*...CAG]atcgctccatgcctta
.....[.....]
.....C...act.....a.....[.....]

(11) (12) (13)
aaaataagcaatatatggaaaatttataagaaccagttttacaaatattgtataagtattcaaaaatgtatcatccagaagaccctatagcattttgcca
.....taa.....C.....C.....t.....
.....g.....g.....a.....a.....C...t.....tt.

tctcatatggaagcatactgaactgagaactctgaattataaccattcagaaaatacttagcacttggaacaaaggaccccaaatggctacttttactc
.....C.....
.....tt.....a.....C.....

(14)
tctgagaagtgaatattgtattcttgc
.....tga.....
.....tga.....C.....

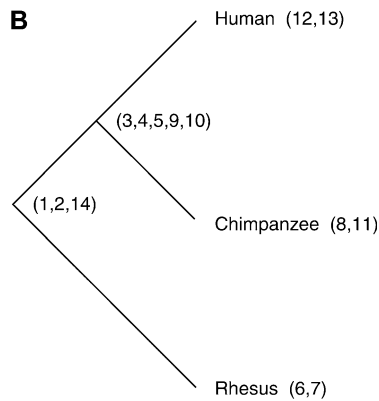


FIG. 3.—Continued.

human $\psi FXR\beta$. Three highly conserved sequence segments, each ~ 15 bp, were found within ~ 250 bp immediately upstream to the coding sequence of $\psi FXR\beta$ (fig. 5A). Further upstream $\sim 4,500$ bp away in an expansive (75 kb) intergenic region between *SIKE* and *SYCP1* resides a ~ 250 -bp sequence segment that is highly conserved across vertebrates between human and chicken (fig. 5B). This sequence segment has a high regulatory potential (>0.35 , see King et al. [2005]), and its mouse orthologous sequence is only 100 bp upstream to the first (noncoding) exon of the mouse *FXR\beta*.

Some NR Pseudogenes Were Derived from Semiprocessed RNA Transcripts

Most retropseudogenes were created from processed RNA transcripts. In this study, however, we found that 2 mouse NR pseudogenes contain remnant introns, which suggests that they were derived from semiprocessed RNA transcripts instead. Mouse $\psi Rev-erb\beta$ on chromosome 19 is such a “semiprocessed pseudogene,” as the fifth of 7 introns of *Rev-erb\beta* was largely retained (fig. 6A). Although its splicing sites remain largely intact, this intron of $\psi Rev-erb\beta$, containing 1,962 nt, is two-thirds of its homologous sequence in *Rev-erb\beta*. In addition to the length difference, these 2 introns share some sequence homology, mainly in their first 500 bases. A closer look also revealed another informative divergence: Although there is no interspersed repeat sequence present in the fifth intron of *Rev-erb\beta*, the intron of $\psi Rev-erb\beta$ hosts 2 short interspersed nuclear elements and 1 long interspersed nuclear element (LINE).

There are 2 $\psi LRH1$ in the mouse genome. Unlike $\psi LRH1$ on chromosome 6, which is a processed pseudo-

gene, $\psi LRH1$ on chromosome 3 has a small intron of 86 bp long in its sequence (fig. 6B). Sequence alignment located this intron at the same place as the third intron, which is over 3.5 kb long, in the coding sequence of *LRH1*. Although 2 introns are greatly different in length, some limited sequence similarity is shared between them, which, in addition to their identical locations in respective genes, suggests that the former originated from the latter and was shortened subsequently. However, the presence of both the additional 3 bases, ATT, before the donor site (GT) and the 24 bases that could not be found in the corresponding intron of *LRH1* is yet to be explained.

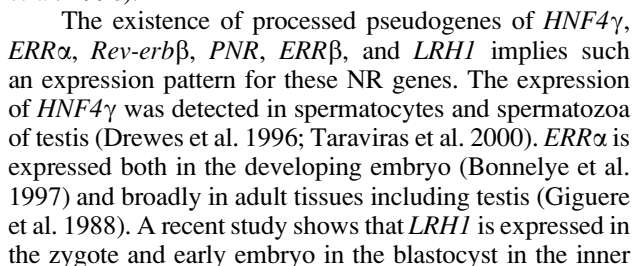
Discussion

NR Pseudogenes Are Scarce

Overall, there are only a very small number of NR pseudogenes in each of the vertebrate genomes examined. Surprisingly, we could not identify any duplicated NR pseudogenes. The absence of duplicated NR pseudogenes is highly unusual because the NR family was expanded through 2 rounds of gene duplications to recognize more ligands as environmental signals: one that gave rise to the various groups of receptors before the arthropod/vertebrate split and the vertebrate-specific one that diversified the constituents of each group by creating the paralogous versions of the various receptors (Laudet 1997). Compared with the human olfactory receptor family, which was expanded through recent gene duplications but contains 359 (53%) duplicated pseudogenes (Glusman et al. 2001), the absence of NR-duplicated pseudogenes suggests that the duplications of the ancestral NR genes were tightly controlled: All NR genes newly created by duplication could successfully subfunctionalize and subsequently evolve into functionally different NR genes.

The number of processed NR pseudogenes is also unexpectedly small. In the human genome, $\sim 8,000$ processed pseudogenes, which originate from $\sim 2,500$ distinct functional genes, have been identified (Zhang et al. 2003)—that is, 3 processed pseudogenes for each functional gene that has been retrotransposed, an average well above that of NR family observed here. Given the size of the NR family (48 in human, 48 expected in chimpanzee, 49 in mouse, and 49 in rat were found in a genome-wide survey, see Zhang et al. [2004]), the scarcity of NR retropseudogenes is further evinced by the comparison with the ribosomal protein-coding genes, which have more than 1,700 (Zhang et al. 2002) retropseudogenes. The scarcity of NR retropseudogenes reflects the overall low expression level and often-times restricted expression locale of the NR genes and could be a general feature of most transcription factor-coding genes.

FIG. 3.—Human, chimpanzee, and rhesus $\psi FXR\beta$. (A) Disruptive defects in $\psi FXR\beta$. Such sequence defects, including frameshifts, nonsense mutations, and splice site mutations, were found in the sequence alignment at 14 orthologous positions, which are numbered and accented in bold underlined letters. For clarity, the base letters in chimpanzee and rhesus $\psi FXR\beta$ sequences identical to their corresponding ones in human $\psi FXR\beta$ were replaced with dots. In this sequence alignment, “[]” marks the intron boundaries, “—” represents the gaps, and “~” the lost orthologous sequences. (B) Lineage specificity of disruptive defects in $\psi FXR\beta$. Defects specific to human, chimpanzee, and rhesus are shown at the corresponding leaf nodes. Defects occurred in an ancestor, shown at a branching node, are found in all its descendants. Thus, defects 1, 2, and 14 are found in all 3 primate species, whereas defects 3, 4, 5, 9, and 10 are found in both human and chimpanzee but not in rhesus.



cell mass, which at gastrulation gives rise, in part, to the germ line (Pare et al. 2004). Although expression of *Rev-erb β* and *PNR* in germ line and early embryo has not been reported, their processed pseudogenes strongly suggest such an expression pattern.

The creation of *FXR* β exemplifies an episode in the second series of duplication events that created the paralogous versions of various receptors in vertebrates (Laudet 1997). Unlike most other paralogous NR genes, however, *FXR* and *FXR* β have been evolving very differently in mammals: *FXR* β is evolving much faster than *FXR* in mammals, but a similar difference in the evolution speed is not observed in nonmammalian vertebrates. It is known that both *FXR* and *FXR* β regulate the biosynthesis of cholesterol (Goodwin et al. 2000; Lu et al. 2000; Otte et al.

The inheritance and fixation of processed pseudogenes in a genome require—as a necessary condition—gene expression in the germ line or cells of the early embryo that contribute to the germ line. It has been shown that the required reverse transcription machinery can be provided by LINEs (Esnault et al. 2000). In addition, endogenous retroviruses (ERVs) can also contribute to the creation of processed pseudogenes (Jamain et al. 2001) as several ERV families are predominantly expressed in germ cells (especially in male germ cells) and in embryonic tissues (Lower et al. 1996).

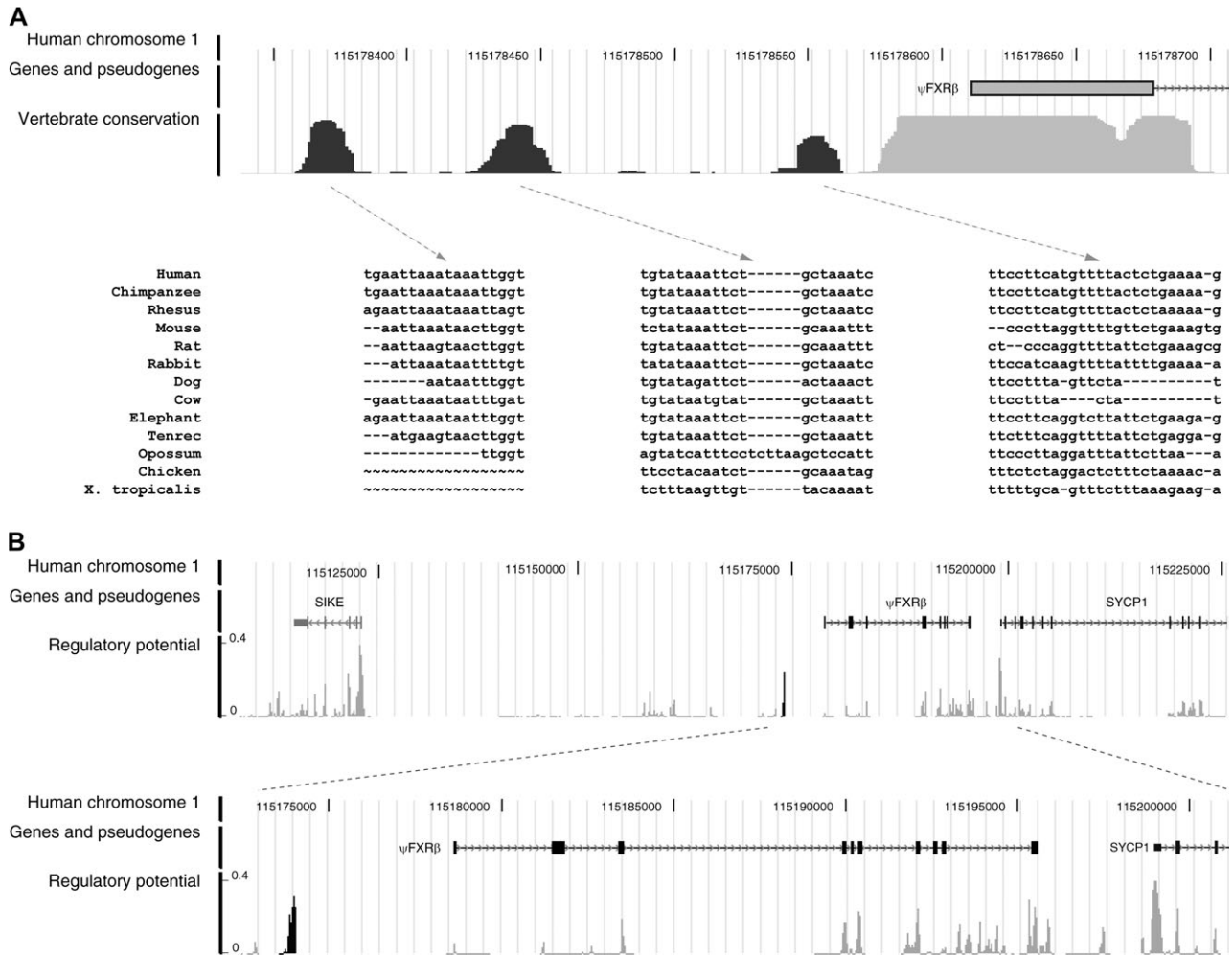
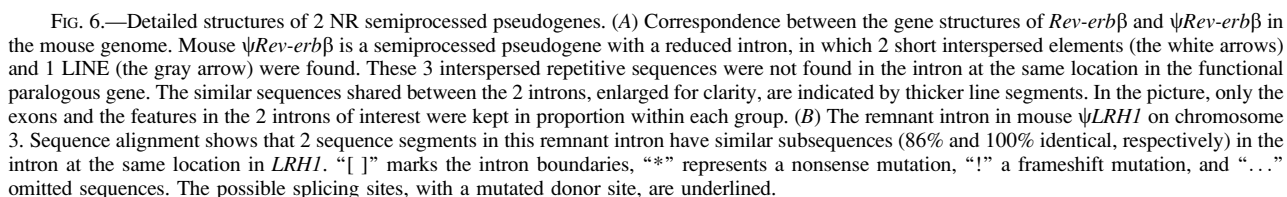


Fig. 5.—Conservation of intergenic sequence upstream to human $\psi FXR\beta$. (A) Three highly conserved sequence segments immediately upstream to the coding sequence of $\psi FXR\beta$ and the alignment of orthologous sequences from 13 vertebrates in these 3 sequence segments. (B) A highly conserved ~250-bp sequence segment with a high regulatory potential 4.5 kb upstream to $\psi FXR\beta$ and a zoom-in view (C). Notice that this sequence segment has a high regulatory potential comparable to that of the transcription start site of the functional gene *SYCP1*.

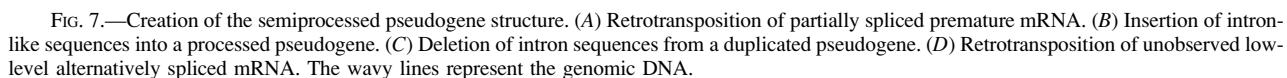
2003). The accelerated evolution, a phenomenon also observed in many other new genes (Begun 1997; Johnson et al. 2001; Maston and Ruvolo 2002; Wang et al. 2002), is needed for FXR β to be subfunctionalized as a receptor for lanosterol, a ligand different from the bile acids, which activates FXR.

Nonfunctionalization of *FXR β* was a relatively recent event. Otte et al. studied *FXR β* in human chimpanzee, gorilla, orangutan, and rhesus monkey, which are all Old World primates and found in all of them the telltale pseudogene defects similar to those in the human ortholog but not in the gene sequences from any other mammals. The date of the *FXR β* silencing based on our calculation indicates that this event postdated the separation of catarrhines and platyrrhines in the primate phylogeny and thus suggests that *FXR β* is not a pseudogene in the New World monkeys, such as marmosets and squirrel monkeys. Given the long evolution of ~ 496 Myr duration since its creation, prior to the nonfunctionalization, *FXR β* had probably already evolved to encode an NR different from *FXR*.

Because the loss of a single-copy gene is usually deleterious and unlikely to be fixed in a population, it remains unclear under what circumstances *FXR β* was silenced—making it an exceedingly rare unitary pseudogene—and how its loss was tolerated and fixed in the ancestral anthropoid population. Two explanations, however, are possible. If the function that *FXR β* provided became redundant in the ancient anthropoids under certain conditions, then ψ *FXR β* could be fixed in the population by random genetic drift under the same conditions because the loss of the *FXR β* product did not constitute a disadvantage, and thus, the selection against the loss was rather weak. This release from selective pressure is believed to be how the nonfunctionalization of L-gulonolactone oxidase could be fixed in humans and guinea pigs (Koshizaka et al. 1988): It has been hypothesized that the guinea pig and human ancestors subsisted on a naturally ascorbic acid-rich diet, and therefore, the loss of the enzyme did not constitute a disadvantage. On the other hand, instead of being a neutral event, the silencing of *FXR β* could be advantageous to the anthropoid ancestors and consequently swept through the population to



It is rather surprising to find $\psi FXR\beta$ to be still transcribed in human even tens of millions of years after its pseudogenization. However, as recent studies have shown, transcription from pseudogenes may be a widely spread cellular phenomenon (Harrison et al. 2005; Zheng et al. 2005, 2007). Just like the transcription of functional genes, the transcription of pseudogenes should also be initiated from their promoters and possibly regulated by other sequence elements as they are transcribed by the same nuclear machinery. However, such *cis*-regulatory elements for pseudogenes have not been reported. The conserved noncoding sequences that we identified with high regulatory potential



upstream to human $\psi FXR\beta$ are possibly such “cryptic” promoter and other functional *cis* elements initiating and regulating its transcription. The conservation of short regulatory *cis* elements, which enables the transcription of pseudogenes long after their nonfunctionalization, may imply that the transcribed pseudogenes and their regulatory *cis* elements together are under negative selection. This in turn suggests that the pseudogene transcripts may play certain functional roles.

Semiprocessed Pseudogenes Provide Insights into the RNA Splicing Process

A retropseudogene is a nonfunctionalized retrosequence, which is generated through a multistep biological process: The DNA is transcribed into pre-mRNA and then processed into mRNA; the mRNA is reverse transcribed into cDNA, which becomes integrated into the genomic DNA. Most retropseudogenes were derived from (fully) processed RNA transcripts, including ones derived from alternatively spliced transcripts (Shemesh et al. 2006), but in rare cases, retropseudogenes such as the mouse $\psi Rev-erb\beta$ and $\psi LRHI$ found in this study were derived from semiprocessed RNA transcripts.

It is conceivable that the semiprocessed pseudogene structure found in a genome could be generated through several different biological processes (fig. 7). Pseudogenes with (remnant) “introns” can be genuine semiprocessed pseudogenes generated from partially spliced premature mRNA (fig. 7A). Such pseudogene structure could also be created by sequence insertion (fig. 7B) or deletion (fig. 7C); however, this is unlikely as the sequence alteration must be highly precise. A processed retropseudogene generated from the unobserved low-level alternatively spliced mRNA (fig. 7D) could also appear as a semiprocessed pseudogene at the first glance when compared with the known mRNA sequence. Sequence insertion could be slightly more probable than the latter 2 processes as intron insertion at the splice site—intron gain—has been observed before (Roy and Gilbert 2006). Nevertheless, the exceedingly low probability for the latter 3 pseudogene generation processes to occur and the sequence characteristics observed in mouse $\psi Rev-erb\beta$ and $\psi LRHI$ argue favorably, if not exclusively, that these 2 pseudogenes are rare semiprocessed retropseudogenes.

By the nature of the generating process, retrosequences should lose their function right at their creation. However, the murine preproinsulin I gene, a functional semiprocessed retrogene, is a rare, if not the sole, exception. In our study, we found no substantial sequence similarity between the regions (up to 5 kb) upstream from the “coding regions” of $\psi Rev-erb\beta$ and *Rev-erb* β in mouse, which suggests that, unlike the murine preproinsulin I retrogene, $\psi Rev-erb\beta$ did not carry any of the *Rev-erb* β promoter and regulatory sequences and thus was silenced on the spot after its retrotransposition. The simultaneity of the duplication and the nonfunctionalization of $\psi Rev-erb\beta$, which freed its coding sequence from selective pressure immediately after retrotransposition, accounts for the similar sequence divergence in all its regions homologous to *Rev-erb* β .

After being transcribed from the DNA, the primary transcripts undergo RNA splicing, a series of processing reactions mediated by the spliceosome to remove the intronic segments. The existence of the semiprocessed pseudogenes signifies that the removal of introns is not a nonstop process proceeding from the start to the end. Instead, it is a collection of discrete splicing events: Each intron is removed by a spliceosome assembled at its splicing sites. This discreteness makes it possible for a semiprocessed pre-mRNA to be “hijacked” and reversely transcribed into cDNAs. However, given the rarity of the semiprocessed pseudogenes, despite being a discrete process, RNA splicing should be a sequence of very fast and efficient removals of all introns from primary RNA transcripts.

Conclusions

We surveyed the NR pseudogenes in 8 vertebrate species whose complete genome sequences are currently available and provide a detailed study of NR pseudogenes in human, chimpanzee, mouse, and rat, giving a complete catalog of their locations, sequences, and defects. In contrast to some highly expressed gene families, such as ones encoding ribosomal proteins and olfactory receptors, NR pseudogenes are scarce in all surveyed genomes, reflecting the temporally and spatially restricted expression pattern of transcription factor–coding genes.

In striking opposition to the initial expectations derived from the mechanisms for pseudogene generation and previous large-scale pseudogene analysis, all but one NR pseudogenes identified in this study are retropseudogenes, and no duplicated NR pseudogenes are found. Through detailed sequence analysis of $\psi FXR\beta$, a previously identified unitary pseudogene in the Old World primates, we could both date its nonfunctionalization in the anthropoid lineage and identify the mutations that most likely caused its silencing. Comparing the noncoding sequence upstream to $\psi FXR\beta$ in human with the orthologous sequences in other vertebrate genomes, we found conserved sequence segments with high regulatory potential. Such short sequences are cryptic *cis*-regulatory elements, as they enable $\psi FXR\beta$, a human pseudogene, to be transcribed. Moreover, gene structure analysis revealed that 2 mouse NR pseudogenes contain remnant introns, which suggests that unlike processed pseudogenes they were derived from semiprocessed RNA transcripts. The finding of such rare semiprocessed pseudogenes indicates that RNA splicing is a sequence of fast and efficient but discrete removals of introns from primary RNA transcripts.

Supplementary Material

Supplementary tables 1 and 2 and figures 1 and 2 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

Z.D.Z. thanks Deyou Zheng for helpful discussion. Z.D.Z. was funded by a National Institutes of Health

(NIH) grant (T15 LM07056) from the National Library of Medicine. This work was supported by grants from NIH/ National Human Genome Research Institute to G.W. and M.G.

Literature Cited

- Adams MD, Celniker SE, Holt RA, et al. (190 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. *Science*. 287:2185–2195.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science*. 297:1003–1007.
- Begun DJ. 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*. *Genetics*. 145:375–382.
- Bonnelye E, Vanacker JM, Spruyt N, Alric S, Fournier B, Desbiens X, Laudet V. 1997. Expression of the estrogen-related receptor 1 (ERR-1) orphan receptor during mouse development. *Mech Dev*. 65:71–85.
- Cheng Z, Ventura M, She X, et al. (12 co-authors). 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*. 437:88–93.
- The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437:69–87.
- Chou HH, Hayakawa T, Diaz S, Krings M, Indriati E, Leakey M, Paabo S, Satta Y, Takahata N, Varki A. 2002. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci USA*. 99:11736–11741.
- Dehal P, Satou Y, Campbell RK, et al. (87 co-authors). 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*. 298:2157–2167.
- Drewes T, Senkel S, Holewa B, Ryffel GU. 1996. Human hepatocyte nuclear factor 4 isoforms are encoded by distinct and differentially expressed genes. *Mol Cell Biol*. 16:925–931.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.
- Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet*. 24:363–367.
- Galili U, Swanson K. 1991. Gene sequences suggest inactivation of alpha-1,3-galactosyltransferase in catarrhines after the divergence of apes from monkeys. *Proc Natl Acad Sci USA*. 88:7401–7404.
- Gibbs RAGM, Weinstock ML, Metzker DM, et al. (229 co-authors). 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 428:493–521.
- Giguere V, Yang N, Segui P, Evans RM. 1988. Identification of a new class of steroid hormone receptors. *Nature*. 331:91–94.
- Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. *Genome Res*. 11:685–702.
- Goodwin B, Jones SA, Price RR, et al. (13 co-authors). 2000. A regulatory cascade of the nuclear receptors FXR, SHP-1, and LRH-1 represses bile acid biosynthesis. *Mol Cell*. 6: 517–526.
- Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res*. 33:2374–2383.
- Hedges SB. 2002. The origin and evolution of model organisms. *Nat Rev Genet*. 3:838–849.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*. 432:695–716.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature*. 431:931–945.
- Jamain S, Gironnot M, Leroy P, Clergue M, Quach H, Fellous M, Bourgeron T. 2001. Transduction of the human gene FAM8A1 by endogenous retrovirus during primate evolution. *Genomics*. 78:38–45.
- Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature*. 413:514–519.
- King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*. 15: 1051–1060.
- Koshizaka T, Nishikimi M, Ozawa T, Yagi K. 1988. Isolation and sequence analysis of a complementary DNA encoding rat liver L-gulonogamma-lactone oxidase, a key enzyme for L-ascorbic acid biosynthesis. *J Biol Chem*. 263: 1619–1621.
- Laudet V. 1997. Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. *J Mol Endocrinol*. 19:207–226.
- Li WH. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.
- Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature*. 292:237–239.
- Lindblad-Toh KC, Wade M, Mikkelsen TS, et al. (236 co-authors). 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 438: 803–819.
- Lower R, Lower J, Kurth R. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci USA*. 93: 5177–5184.
- Lu TT, Makishima M, Repa JJ, Schoonjans K, Kerr TA, Auwerx J, Mangelsdorf DJ. 2000. Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors. *Mol Cell*. 6:507–515.
- Maglich JM, Sluder A, Guan X, Shi Y, McKee DD, Carrick K, Kamdar K, Willson TM, Moore JT. 2001. Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila* genomes. *Genome Biol*. 2: RESEARCH0029.
- Maston GA, Ruvolo M. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol Biol Evol*. 19:320–335.
- Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N. 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol*. 4:R74.
- Otte K, Kranz H, Kober I, et al. (18 co-authors). 2003. Identification of farnesoid X receptor beta as a novel mammalian nuclear receptor sensing lanosterol. *Mol Cell Biol*. 23:864–872.
- Pare JF, Malenfant D, Courtemanche C, Jacob-Wagner M, Roy S, Allard D, Belanger L. 2004. The fetoprotein transcription factor (FTF) gene is essential to embryogenesis and cholesterol homeostasis and is regulated by a DR4 element. *J Biol Chem*. 279:21206–21216.

- Robinson-Rechavi M, Carpentier AS, Duffraisse M, Laudet V. 2001. How many nuclear hormone receptors are there in the human genome? *Trends Genet.* 17:554–556.
- Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet.* 7: 211–221.
- Shemesh R, Novik A, Edelheit S, Sorek R. 2006. Genomic fossils as a snapshot of the human transcriptome. *Proc Natl Acad Sci USA.* 103:1364–1369.
- Sladek R, Beatty B, Squire J, Copeland NG, Gilbert DJ, Jenkins NA, Giguere V. 1997. Chromosomal mapping of the human and murine orphan receptors ERRalpha (ESRRA) and ERRbeta (ESRRB) and identification of a novel human ERRalpha-related pseudogene. *Genomics.* 45:320–326.
- Sluder AE, Mathews SW, Hough D, Yin VP, Maina CV. 1999. The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Res.* 9:103–120.
- Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature.* 428: 415–418.
- Taraviras S, Mantamadiotis T, Dong-Si T, Mincheva A, Lichter P, Drewes T, Ryffel GU, Monaghan AP, Schutz G. 2000. Primary structure, chromosomal mapping, expression and transcriptional activity of murine hepatocyte nuclear factor 4gamma. *Biochim Biophys Acta.* 1490:21–32.
- Tchenio T, Segal-Bendirdjian E, Heidmann T. 1993. Generation of processed pseudogenes in murine cells. *EMBO J.* 12: 1487–1497.
- Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. *Genome Res.* 13: 2559–2567.
- Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc Natl Acad Sci USA.* 99:4448–4453.
- Waterston RHK, Lindblad-Toh E, Birney J, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 420:520–562.
- Zhang Z, Burch PE, Cooney AJ, Lanz RB, Pereira FA, Wu J, Gibbs RA, Weinstock G, Wheeler DA. 2004. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res.* 14:580–590.
- Zhang Z, Gerstein M. 2003. The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene.* 312:61–72.
- Zhang Z, Harrison P, Gerstein M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* 12:1466–1482.
- Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 13:2541–2558.
- Zheng D, Frankish A, Baertsch R, et al. (16 co-authors) Forthcoming. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription and evolution. *Genome Res.* 17:839–851.
- Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, Gerstein M. 2005. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol.* 349:27–45.

Dan Graur, Associate Editor

Accepted October 25, 2007