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, \( \psi_{\text{Rev-erb}} \) and \( \psi_{\text{LRH1}} \). 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, \( \text{FXR}^b \) 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 \( \text{FXR}^a–\text{FXR}^b \) duplication and the nonfunctionalization of \( \text{FXR}^b \) 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, morphogenesis, 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 (\( \psi \)) 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: \( \psi_{\text{ERR}} \) (Sladek et al. 1997), \( \psi_{\text{HNF}} \gamma \) (Tcherni et al. 1993), and \( \psi_{\text{FXR}}^b \) (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; Torres 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
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 which these predictions were made. Alignments were then used to estimate the date of the divergence of FXR and FXRβ by using the Neighbor-Joining algorithm in the MUSCLE (Edgar 2004).

### 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—ψERRα (Sladek et al. 1997), ψHNF4γ (Tchenio et al. 1993), and ψFXRβ (Majlich et al. 2001; Otte et al. 2003)—was confirmed by our analysis. Except for 1 human NR pseudogene, ψFXRβ, 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 translation factors and generally have low and restricted transcription profiles, we expected most of the NR pseudogenes to be created by duplication.

Two ψERRα Are in the Human Genome

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

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Symbols of NR Used in the Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Official Name</td>
</tr>
<tr>
<td>FXRβ</td>
<td>NR1H5</td>
</tr>
<tr>
<td>HNF4γ</td>
<td>NR2A2</td>
</tr>
<tr>
<td>ERRα</td>
<td>NR3B1</td>
</tr>
<tr>
<td>Rev-erbβ</td>
<td>NR1D2</td>
</tr>
<tr>
<td>PNR</td>
<td>NR2E3</td>
</tr>
<tr>
<td>ERRβ</td>
<td>NR3B2</td>
</tr>
<tr>
<td>LRH1</td>
<td>NR5A2</td>
</tr>
</tbody>
</table>

### Table 2

Symbols of NR Used in the Text

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Official Name</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>FXRβ</td>
<td>NR1H5</td>
<td>Farnesoid X receptor, beta</td>
</tr>
<tr>
<td>HNF4γ</td>
<td>NR2A2</td>
<td>Hepatocyte nuclear factor 4, gamma</td>
</tr>
<tr>
<td>ERRα</td>
<td>NR3B1</td>
<td>Estrogen-related receptor, alpha</td>
</tr>
<tr>
<td>Rev-erbβ</td>
<td>NR1D2</td>
<td>Thyroid hormone receptor, alpha-like</td>
</tr>
<tr>
<td>PNR</td>
<td>NR2E3</td>
<td>Photoreceptor-specific NR</td>
</tr>
<tr>
<td>ERRβ</td>
<td>NR3B2</td>
<td>Estrogen-related receptor, beta</td>
</tr>
<tr>
<td>LRH1</td>
<td>NR5A2</td>
<td>Liver receptor homolog 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Official Name</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>FXRβ</td>
<td>NR1H5</td>
<td>Farnesoid X receptor, beta</td>
</tr>
<tr>
<td>HNF4γ</td>
<td>NR2A2</td>
<td>Hepatocyte nuclear factor 4, gamma</td>
</tr>
<tr>
<td>ERRα</td>
<td>NR3B1</td>
<td>Estrogen-related receptor, alpha</td>
</tr>
<tr>
<td>Rev-erbβ</td>
<td>NR1D2</td>
<td>Thyroid hormone receptor, alpha-like</td>
</tr>
<tr>
<td>PNR</td>
<td>NR2E3</td>
<td>Photoreceptor-specific NR</td>
</tr>
<tr>
<td>ERRβ</td>
<td>NR3B2</td>
<td>Estrogen-related receptor, beta</td>
</tr>
<tr>
<td>LRH1</td>
<td>NR5A2</td>
<td>Liver receptor homolog 1</td>
</tr>
</tbody>
</table>

The tree was rooted by LXRα.
Table 2  
Human and Rodent NR Pseudogenes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Pseudogene</th>
<th>Accession</th>
<th>Location</th>
<th>Truncation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>(\psi{\text{FXR}}\beta)</td>
<td>15259</td>
<td>1p13.1+</td>
<td>Unitary</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{HNF}}4\gamma)</td>
<td>128390</td>
<td>13q21.1-</td>
<td>Unitary</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{ERR}}\lambda)</td>
<td>5136</td>
<td>13q12.11-</td>
<td>Processed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{ERR}}\zeta)</td>
<td>24162</td>
<td>13q12.11+</td>
<td>Processed</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>(\psi{\text{FXR}}\beta)</td>
<td>8400</td>
<td>13-</td>
<td>Unitary</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{HNF}}4\gamma)</td>
<td>8401</td>
<td>13-</td>
<td>Unitary</td>
</tr>
<tr>
<td>Mouse</td>
<td>(\psi{\text{REV-ERB}}\beta)</td>
<td>19393</td>
<td>9q33+</td>
<td>Semiprocessed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{PXR}})</td>
<td>6324</td>
<td>15q23.1+</td>
<td>Processed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{ERR}}\gamma)</td>
<td>10804</td>
<td>XqA5+</td>
<td>Processed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{LHRH}})</td>
<td>8260</td>
<td>3qH2+</td>
<td>Semiprocessed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{LHRH}})</td>
<td>17110</td>
<td>6qF1-</td>
<td>Processed</td>
</tr>
<tr>
<td>Rat</td>
<td>(\psi{\text{FXR}}\beta)</td>
<td>8720</td>
<td>X3q36+</td>
<td>Processed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{LHRH}})</td>
<td>1916</td>
<td>11q21+</td>
<td>Processed</td>
</tr>
<tr>
<td></td>
<td>(\psi{\text{LHRH}})</td>
<td>17561</td>
<td>X1q14-</td>
<td>Processed</td>
</tr>
</tbody>
</table>


\* 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.

\* 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 \(\psi{\text{ERR}}\zeta\) (\(\psi{\text{ERR}}\lambda +\) and \(\psi{\text{ERR}}\lambda -\)), 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 \(\psi{\text{ERR}}\lambda\), 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 pseudogenes in their study.

These 2 human \(\psi{\text{ERR}}\lambda\) 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 \(\psi{\text{ERR}}\zeta\) on the forward strand contains 5 frameshifts, the \(\psi{\text{ERR}}\lambda\) 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 \(\psi{\text{ERR}}\lambda\) are of full length and highly similar, albeit defunct, copies of the transcript of the functional gene, which suggests a young age (~38 MYA), human-specific segmental duplication (Bailey et al. 2002; Cheng et al. 2005). Thus, human \(\psi{\text{ERR}}\lambda +\) 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 \(\psi{\text{FXR}}\beta\) Is a Unitary Pseudogene with Multiple Nonfunctionalization Mutations

Previous studies (Maglich et al. 2001; Otte et al. 2003) have shown that human \(\text{FXR}\beta\) 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 \(\text{FXR}\beta\) protein sequence to the 3-frame translation of the human genomic sequence reveals that the “coding sequence” of the original human \(\text{FXR}\beta\) gene was 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 \( \psi FXR \)β. The mouse FXRβ protein sequence and the translation of the human genomic sequence at the \( \psi FXR \)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 \( \psi FXR \)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).

**Legend**
- ▼ Nonfunctionalization site
- ▼ Frame shift
- ▼ Nonstop mutation
- ▼ Splice site mutation
- ▼ Intron boundaries
Human $\psi$FXR$\beta$ 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$\beta$ 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 $\psi$FXR$\beta$ from these 3 primates (fig. 3B). The disruptive mutations at the 1st, 2nd, and 14th positions in $\psi$FXR$\beta$ 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$\beta$ 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$\beta$ 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$\beta$ in the genomes of representative 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$\beta$ reveals that by the measure of branch length (data not shown), FXR$\beta$ 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$\beta$ sequences, our calculation indicates that the silencing of FXR$\beta$ happened ~42 MYA.

**Intergenic Sequences Immediately Upstream and Downstream to Human $\psi$FXR$\beta$ Are Conserved**

Human $\psi$FXR$\beta$ 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 $\psi$FXR$\beta$ have remained largely intact and functional even long after the inactivation of $\psi$FXR$\beta$. 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
human $\psi FXR\beta$. Three highly conserved sequence segments, each $\sim 15$ bp, were found within $\sim 250$ 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 $\sim 250$-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 pseudogene, $\psi LRH1$ 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.
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).

The existence of processed pseudogenes of \textit{HNF4c}, \textit{ERRa}, \textit{Rev-erb}b, \textit{PNR}, \textit{ERR}b, and \textit{LRH1} implies such an expression pattern for these NR genes. The expression of \textit{HNF4c} was detected in spermatocytes and spermatozoa of testis (Drewes et al. 1996; Taraviras et al. 2000). \textit{ERRa} is expressed both in the developing embryo (Bonnellye et al. 1997) and broadly in adult tissues including testis (Giguere et al. 1988). A recent study shows that \textit{LRH1} is expressed in the zygote and early embryo in the blastocyst in the inner cell mass, which at gastrulation gives rise, in part, to the germ line (Pare et al. 2004). Although expression of \textit{Rev-erb}b and \textit{PNR} in germ line and early embryo has not been reported, their processed pseudogenes strongly suggest such an expression pattern.

Nonfunctionalization of \textit{FXR}b Was a Rare Event that Happened in the Evolution of Anthropoids

The creation of \textit{FXR}b 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, \textit{FXR} and \textit{FXR}b in mammals: \textit{FXR}b is evolving much faster than \textit{FXR} in mammals, but a similar difference in the evolution speed is not observed in nonmammalian vertebrates. It is known that both \textit{FXR} and \textit{FXR}b regulate the biosynthesis of cholesterol (Goodwin et al. 2000; Lu et al. 2000; Otte et al. 2000).
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 FXRβ to be subfunctionalized as a receptor for lanosterol, a ligand different from the bile acids, which activates FXR. Nonfunctionalization of FXRβ was a relatively recent event. Otte et al. studied FXRβ 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 FXRβ silencing based on our calculation indicates that this event postdated the separation of catarrhines and platyrrhines in the primate phylogeny and thus suggests that FXRβ 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, FXRβ 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 FXRβ was silenced—making it an exceeding 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 FXRβ provided became redundant in the ancient anthropoids under certain conditions, then FXRβ could be fixed in the population by random genetic drift under the same conditions because the loss of the FXRβ 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-gulono-γ-lactone 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 FXRβ could be advantageous to the anthropoid ancestors and consequently swept through the population to
fixation—the kind of adaptive evolution illustrated by the inactivation of the α-1,3-galactosyltransferase gene in catarrhines (Galili and Swanson 1991) and the sarcomeric myosin gene (Stedman et al. 2004) and the CMP-N-acetylneuraminic acid hydroxylase gene (Chou et al. 2002) in humans as there seems to be a correlation between pseudogenization and physiological/anatomic changes. To our knowledge, no such correlation has been investigated for FXRβ inactivation. Until more data become available and further analyses are carried out, it remains unclear what was the fixation route—random genetic drift or positive selection—of FXRβ.

It is rather surprising to find FXRβ 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

![Diagram](https://example.com/diagram.png)

**Fig. 6.**—Detailed structures of 2 NR semiprocessed pseudogenes. (A) Correspondence between the gene structures of Rev-erbβ and ψRev-erbβ in the mouse genome. Mouse ψRev-erbβ is a semiprocessed pseudogene with a reduced intron, in which 2 short interspersed elements (the white arrows) and 1 LINE (the gray arrow) were found. These 3 interspersed repetitive sequences were not found in the intron at the same location in the functional paralogous gene. The similar sequences shared between the 2 introns, enlarged for clarity, are indicated by thicker line segments. In the picture, only the exons and the features in the 2 introns of interest were kept in proportion within each group. (B) The remnant intron in mouse ψLRH1 on chromosome 3. Sequence alignment shows that 2 sequence segments in this remnant intron have similar subsequences (86% and 100% identical, respectively) in the intron at the same location in LRH1. “|” marks the intron boundaries. “*” represents a nonsense mutation, “!” a frameshift mutation, and “...” omitted sequences. The possible splicing sites, with a mutated donor site, are underlined.

**Fig. 7.**—Creation of the semiprocessed pseudogene structure. (A) Retrotransposition of partially spliced premature mRNA. (B) Insertion of intron-like sequences into a processed pseudogene. (C) Deletion of intron sequences from a duplicated pseudogene. (D) Retrotransposition of unobserved low-level alternatively spliced mRNA. The wavy lines represent the genomic DNA.
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 $\psi_{LRH1}$ 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 $\psi_{LRH1}$ argue favorably, if not exclusively, that these 2 pseudogenes are rare semiprocessed retropseudogenes.

By the nature of the generating process, retropseudogenes should lose their function right at their creation. However, the murine preproinsulin I gene, a functional semiprocessed retrogene, is 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 $\psi_{Rev-erb}\beta$ in mouse, which suggests that, unlike the murine preproinsulin I retrogene, $\psi_{Rev-erb}\beta$ did not carry any of the Rev-erb$\beta$ 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$.\beta$.

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


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.


Dan Graur, Associate Editor

Accepted October 25, 2007