Analysis of nuclear receptor pseudogenes in vertebrates: How the silent tell their stories

Zhengdong D. Zhang [1], Philip Cayting [1], George Weinstock [4], Mark Gerstein [1,2,3,§]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA; [2] Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA; [3] Department of Computer Science, Yale University, New Haven, CT 06520, USA; [4] Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

[§] Corresponding author (E-mail: zdzmg@bioinfo.mbb.yale.edu)

Running title: Vertebrate Nuclear Receptor Pseudogenes

Keywords: nuclear receptor, pseudogene, nonfunctionalization, protein evolution

+

Abstract

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 eight 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, (i) despite its large size, the NR gene family has only a very small number of pseudogenes in each of the vertebrate genomes examined; (ii) despite the low transcription levels of NR genes, except for one, all other NR pseudogenes identified in this study are retropseudogenes; (iii) 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 two mouse retropseudogenes, *ψRev-erbβ* and *ψLRH1*. Generated from partially processed pre-mRNAs, they appear to be rare examples of highly unusual 'semiprocessed' pseudogenes. Secondly, 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, *FXRβ* 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-FXRβ* duplication and the nonfunctionalization of *FXRβ* in primates.

Background

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 of the DNA-binding domain (DBD) and the ligand-binding domain (LBD), the two 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 melonogaster*

(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, that 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 two 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 three NR pseudogenes in human: *ψERRα* (Sladek et al. 1997), *ψHNF4γ* (Tchenio, Segal-Bendirdjian, and Heidmann 1993), and *ψFXRβ* (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 eight vertebrate genome sequences (Waterston et al. 2002; Gibbs et al. 2004; International Chicken Genome Sequencing Consortium 2004; International Human Genome Sequencing Consortium 2004; Lindblad-Toh et al. 2005; The Chimpanzee Sequencing and Analysis Consortium 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 eight 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, Harrison, and Gerstein 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 expectations. The analysis of these pseudogenes affords unique insights into the evolution and dynamics of this gene family and the mammalian genomes at large.

Results

Nuclear receptor pseudogenes in vertebrate model organisms

By using manual annotation and a pseudogene identification pipeline, we assigned nuclear receptor pseudogenes in human, chimpanzee, mouse, rat, dog, chicken, tetraodon, and zebrafish—eight 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 nuclear receptor pseudogenes in each of the vertebrate genomes examined. Within the human, chimpanzee, mouse, and rat genomes, four, three, five, and three NR pseudogenes were identified respectively (Table 2). The existence of the three previously reported pseudogenes in the human genome—*ψERRα* (Sladek et al. 1997), *ψHNF4γ* (Tchenio, Segal-Bendirdjian, and Heidmann 1993), and *ψFXRβ* (Maglich et al. 2001; Otte et al. 2003)—was confirmed by our analysis. Except for one 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, since NR genes encode transcription factors and generally have low and restricted transcription profiles, we expected most of NR pseudogenes to be created by duplication.

Two ψERRα are in the human genome

Sladek et al. reported the isolation of a processed *ERRα* pseudogene mapped to human chromosome 13q12.1 (Sladek et al. 1997). In our study, however, two processed *ψERRα*s (*ψERRα*+ and *ψERRα*−), immediately next to each other on opposite DNA strands, were identified in the same chromosome band (13q12.11). The genomic sequence interval between these two *ψERRα*, 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 two human *ψERRα* 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 two sequences, is only 3.65% and the number of nucleotide substitution per site between them, K, is 0.038±0.006. The *ψERRα* on the forward strand contains five frame shifts, the *ψERRα* on the reverse strand has four, and both have a premature stop codon at different positions. Of these defects in their sequences, three frame shifts are identical. Except for several internal deletions, both *ψERRα* are 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 *ψERRα+* is absent. This absence indicates that *ψERRα−* 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 *ψERRα+* and *ψERRα−* suggest that the former was created by the duplication of the latter in the human lineage after its divergence from chimpanzee. In fact, those two pseudogenes reside in two expansive (>14.6-kb) and highly similar (96% identical) sequence segments in the human chromosome 13 that were created by a recent(< 6 million years ago), human-specific segmental duplication (Bailey et al. 2002; Cheng et al. 2005). Thus, human *ψERRα+* 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 ψFXRβ is a unitary pseudogene with multiple nonfunctionalization mutations

Previous studies (Maglich et al. 2001; Otte et al. 2003) have shown that human *FXRβ* 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 FXRβ protein sequence to the three-frame translation of the human genomic sequence reveals that the coding sequence of the original human *FXRβ* gene were interrupted by at least nine introns and in the currently defunct gene there are ten disruptive defects, which consist of three frame shifts, four nonsense mutations, and three splice site mutations (Figure 1). These defects are equally distributed at the beginning and the end of this pseudogene.

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

Based on four pairwise comparisons among the mouse and rat *FXR* and *FXRβ* sequences, our study dated the ancient gene duplication event that created this pair of paralogous genes to be ~496 million years ago (Mya) prior to the speciation events (~450 Mya) that ultimately gave rise to fishes and other vertebrates (Figure 4A). This estimation was confirmed by the search result for *FXR* and *FXRβ* in the genomes of representative species that both genes exist in human, chimpanzee, mouse, chicken, frog (*Xenopus tropicalis*), and fish (both zebrafish and pufferfish, Supplementary figure 1). The phylogeny of *FXR* and *FXRβ* reveals that by the measure of branch length (data not shown) *FXRβ* is evolving at least 5.6 times faster than *FXR* in mammals, but a similar difference in the evolution speed is not observed in non-mammalian vertebrates (Figure 4B, see Supplementary figure 2 for the multiple sequence alignment). Based on human, mouse, rat, and dog *FXRβ* sequences, our calculation indicates that the silencing of *FXRβ* happened ~42 Mya,

Intergenic sequences immediately upstream and downstream to human ψFXRβ are conserved

Human *ψFXRβ* is a transcribed pseudogene: real-time quantitative PCR 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 *ψFXRβ* have remained largely intact and functional even long after the inactivation of *ψFXRβ*. 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 *ψFXRβ*. Three highly conserved sequence segments, each ~15 bp, were found within ~250 bp immediately upstream to the 'coding sequence' of *ψFXRβ* (Figure 5A). Further upstream ~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 (Figure 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 *FXRβ*

Some NR pseudogenes were derided from semiprocessed RNA transcripts

Most retropseudogenes were created from processed RNA transcripts. In this study, however, we found two mouse NR pseudogenes contain remnant introns, which suggests that they were derived from semiprocessed RNA transcripts instead. Mouse *ψRev-erbβ* on chromosome 19 is such a 'semiprocessed pseudogene,' as the fifth of seven introns of Rev-erbβ was largely retained (Figure 6A). While its splicing sites remain largely intact, this intron of *ψRev-erbβ*, containing 1962 nucleotides, is two thirds of its homologous sequence in *Rev-erbβ*. In addition to the length difference, these two introns share some sequence homology, mainly in their first 500 bases. A closer look also revealed another informative divergence: while there is no interspersed repeat sequence present in the fifth intron of *Rev-erbβ*, the intron of *ψRev-erbβ* hosts two SINEs and one LINE.

There are two *ψLRH1* in the mouse genome. Unlike *ψLRH1* on chromosome 6, which is a processed pseudogene, *ψLRH1* on chromosome 3 has a small intron of 86 base pairs long in its sequence (Figure 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. While two 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 the former originated from the latter and was shortened subsequently. However, the presence of both the additional three 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 nuclear receptor 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 two 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 vertebrates-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, ~8,000 processed pseudogenes, which originate from ~2,500 distinct functional genes, have been identified (Zhang et al. 2003)—i.e., three 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 reference (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, Harrison, and Gerstein 2002) retropseudogenes. The scarcity of NR retropseudogenes reflects the overall low expression level and oftentimes 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 long interspersed elements (Esnault, Maestre, and Heidmann 2000). In addition, endogenous retroviruses (ERV) 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, Lower, and Kurth 1996).

The existence of processed pseudogenes of *HNF4γ*, *ERRα*, *Rev-erbβ*, *PNR*, *ERRβ*, and *LRH1* implies such an expression pattern for these NR genes. The expression of *HNF4γ* was detected in spermatocytes and spermatozoa of testis (Drewes et al. 1996; Taraviras et al. 2000). *ERRα* is expressed both in the developing embryo (Bonnelye et al. 1997) and broadly in adult tissues including testis (Giguere et al. 1988). A recent study shows that *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 *Rev-erbβ*, *PNR* in germ line and early embryo has not been reported, their processed pseudogenes strongly suggest such an expression pattern.

Nonfunctionalization of FXRβ was a rare event that happened in the evolution of anthropoids

The creation of *FXRβ* 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 *FXRβ* have been evolving very differently in mammals: *FXRβ* is evolving much faster than *FXR* in mammals, but a similar difference in the evolution speed is not observed in non-mammalian vertebrates. It is known that both FXR and FXRβ regulate the biosynthesis of cholesterol (Goodwin et al. 2000; Lu et al. 2000; Otte et al. 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 FXRβ to be subfunctionalized as a receptor for lanosterol, a ligand different from the bile acids, which activate 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 *FXRβ* is not a pseudogene in the New World monkeys, such as marmosets and squirrel monkeys. Given the long evolution of ~496 million years' duration since its creation, prior to the nonfunctionalization, *FXRβ* had probably already evolved to encode a nuclear receptor different from FXR.

Since 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 *ψFXRβ* 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), 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 *ψFXRβ*.

It is rather surprising to find *ψFXRβ* 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; Zheng et al. 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 *ψFXRβ* 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 multi-step 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 *ψRev-erbβ* and *ψLRH1* 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 (Figure 7). Pseudogenes with (remnant) 'introns' can be genuine semiprocessed pseudogenes generated from partially spliced premature mRNA (Figure 7A). Such pseudogene structure could also be created by sequence insertion (Figure 7B) or deletion (Figure 7C), however unlikely as the sequence alteration must be highly precise. A processed retropseudogene generated from the unobserved low-level alternatively spliced mRNA (Figure 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 two 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 three pseudogene generation processes to occur and the sequence characteristics observed in mouse *ψRev-erbβ* and *ψLRH1* argue favorably, if not exclusively, that these two 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 *ψRev-erbβ* and *Rev-erbβ* in mouse, which suggests that, unlike the murine preproinsulin I retrogene, *ψRev-erbβ* did not carry any of the *Rev-erbβ* promoter and regulatory sequences and thus was silenced on the spot after its retrotransposition. The simultaneity of the duplication and the nonfunctionalization of *ψRev-erbβ*, which freed its coding sequence from selective pressure immediately after retrotransposition, accounts for the similar sequence divergence in all its regions homologous to *Rev-erbβ*.

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 non-stop 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 nuclear receptor pseudogenes in eight 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 catalogue 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 *ψFXRβ*, a previously identified unitary pseudogene in the Old World primates, we could both date its nonfucntionalization in the anthropoid lineage and identify the mutations that most likely caused its silencing. Comparing the non-coding sequence upstream to *ψFXRβ* in human with the orthologous sequences in other vertebrate genomes, we found conserved sequence segments with high regulatory potential. Such short sequences could be cryptic promoter and other *cis*-regulatory elements that enable the transcription of *ψFXRβ* observed in human. Moreover, gene structure analysis revealed that two 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.

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 three 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 frame shifts 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-FXRβ* duplication (T_D), four homologous sequences, FXR_{mouse}, $FXR\beta_{mouse}$, FXR_{rat}, and $FXR\beta_{rat}$, were used (Li 1997). Since the synonymous substitutions per synonymous site (*Ks*) are large and thus cannot be estimated accurately, they are not used to calculate T_D. As the equation shows below, only the nonsynonymous substitution per nonsynonymous site (*Ka*) are used. T_D is estimated by

$$T_D = 2 \cdot T_S \cdot \frac{\overline{K}_{a\ FXR,FXR\beta}}{K_{a\ FXR} + K_{a\ FXR\beta}}$$

where T_S is the divergence time between mouse and rat, for which 41 million years were used in the calculation (Hedges 2002), $\overline{K}_{a\ FXR,FXR\beta}$ is the average value of four numbers of nucleotide substitutions per site estimated from four pairwise comparisons: FXR_{mouse}-$FXR\beta_{mouse}$, FXR_{mouse}-$FXR\beta_{rat}$, FXR_{rat}-$FXR\beta_{mouse}$, and FXR_{rat},-$FXR\beta_{rat}$, $K_{a\ FXR}$ and $K_{a\ FXR\beta}$ are the numbers of the synonymous substitutions per synonymous site in *FXR* and *FXRβ* respectively (Supplementary table 1).

To estimate the nonfunctionalization time (T_N) of *ψFXRβ* in the primate lineage, we used the method devised by Chou et al. See the reference (Chou et al. 2002) for a detailed description of the method. Briefly, it assumes that non-synonymous mutations are selected against until the gene is inactivated; thereafter mutations at both synonymous and non-synonymous sites accumulate at the neutral mutation rate. Quantification of lineage-specific mutation rates at synonymous and non-synonymous sites remote from the inactivating deletion provides the information necessary for the calculation. Four *FXRβ* sequences, from human, mouse, rat, and chicken, were used for the calculation (Supplementary table 2). We used the method proposed by Li et al. (Li, Gojobori, and Nei 1981) to estimate the nonfunctionalization time of all retropseudogenes identified in this study. Because they are 'dead on arrival', we assumed that $T_N = T_D$.

Multiple FXR and FXRβ peptide sequences together with the human LXRα peptide sequences were aligned using MUSCLE (Edgar 2004). The phylogeny of *FXR* and *FXRβ* 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 LXRα.

List of abbreviations

DBD	DNA binding domain
ERV	endogenous retroviruses
LBD	ligand binding domain
LINE	long interspersed nuclear elements
NR	nuclear receptor
SINE	short interspersed nuclear elements

Acknowledgments

Z.D.Z. thanks Deyou Zheng for helpful discussion. Z.D.Z. was funded by an NIH grant (T15 LM07056) from the National Library of Medicine. This work was supported by grants from NIH/NHGRI to G.W. and M.G.

References

Adams, M. D.S. E. CelnikerR. A. HoltC. A. EvansJ. D. GocayneP. G. AmanatidesS. E. SchererP. W. LiR. A. HoskinsR. F. GalleR. A. GeorgeS. E. LewisS. RichardsM. AshburnerS. N. HendersonG. G. SuttonJ. R. WortmanM. D. YandellQ. ZhangL. X. ChenR. C. BrandonY. H. RogersR. G. BlazejM. ChampeB. D. PfeifferK. H. WanC. DoyleE. G. BaxterG. HeltC. R. NelsonG. L. GaborJ. F. AbrilA. AgbayaniH. J. AnC. Andrews-PfannkochD. BaldwinR. M. BallewA. BasuJ. BaxendaleL. BayraktarogluE. M. BeasleyK. Y. BeesonP. V. BenosB. P. BermanD. BhandariS. BolshakovD. BorkovaM. R. BotchanJ. BouckP. BroksteinP. BrottierK. C. BurtisD. A. BusamH. ButlerE. CadieuA. CenterI. ChandraJ. M. CherryS. CawleyC. DahlkeL. B. DavenportP. DaviesB. de PablosA. DelcherZ. DengA. D. MaysI. DewS. M. DietzK. DodsonL. E. DoupM. DownesS. Dugan-RochaB. C. DunkovP. DunnK. J. DurbinC. C. EvangelistaC. FerrazS. FerrieraW. FleischmannC. FoslerA. E. GabrielianN. S. GargW. M. GelbartK. GlasserA. GlodekF. GongJ. H. GorrellZ. GuP. GuanM. HarrisN. L. HarrisD. HarveyT. J. HeimanJ. R. HernandezJ. HouckD. HostinK. A. HoustonT. J. HowlandM. H. WeiC. IbegwamM. JalaliF. KalushG. H. KarpenZ. KeJ. A. KennisonK. A. KetchumB. E. KimmelC. D. KodiraC. KraftS. KravitzD. KulpZ. LaiP. LaskoY. LeiA. A. LevitskyJ. LiZ. LiY. LiangX. LinX. LiuB. MatteiT. C. McIntoshM. P. McLeodD. McPhersonG. MerkulovN. V. MilshinaC. MobarryJ. MorrisA. MoshrefiS. M. MountM. MoyB. MurphyL. MurphyD. M. MuznyD. L. NelsonD. R. NelsonK. A. NelsonK. NixonD. R. NusskernJ. M. PaclebM. PalazzoloG. S. PittmanS. PanJ. PollardV. PuriM. G. ReeseK. ReinertK. RemingtonR. D. SaundersF. ScheelerH. ShenB. C. ShueI. Siden-KiamosM. SimpsonM. P. SkupskiT. SmithE. SpierA. C. SpradlingM. StapletonR. StrongE. SunR. SvirskasC. TectorR. TurnerE. VenterA. H. WangX. WangZ. Y. WangD. A. WassarmanG. M. WeinstockJ. WeissenbachS. M. WilliamsWoodageTK. C. WorleyD. WuS. YangQ. A. YaoJ. YeR. F. YehJ. S. ZaveriM. ZhanG. ZhangQ. ZhaoL. ZhengX. H. ZhengF. N. ZhongW. ZhongX. ZhouS. ZhuX. ZhuH. O. SmithR. A. GibbsE. W. MyersG. M. Rubin, and J. C. Venter. 2000. The genome sequence of Drosophila melanogaster. Science **287**:2185-2195.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. 2002. Recent segmental duplications in the human genome. Science **297**:1003-1007.

Begun, D. J. 1997. Origin and evolution of a new gene descended from alcohol dehydrogenase in Drosophila. Genetics **145**:375-382.

Bonnelye, E., J. M. Vanacker, N. Spruyt, S. Alric, B. Fournier, X. Desbiens, and V. Laudet. 1997. Expression of the estrogen-related receptor 1 (ERR-1) orphan receptor during mouse development. Mech Dev **65**:71-85.

Cheng, Z., M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R. K. Wilson, S. Paabo, M. Rocchi, and E. E. Eichler. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature **437**:88-93.

Chou, H. H., T. Hayakawa, S. Diaz, M. Krings, E. Indriati, M. Leakey, S. Paabo, Y. Satta, N. Takahata, and A. Varki. 2002. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc Natl Acad Sci U S A **99**:11736-11741.

Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Necula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H. G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-Bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B. I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D. S. Rokhsar. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science **298**:2157-2167.

Drewes, T., S. Senkel, B. Holewa, and G. U. Ryffel. 1996. Human hepatocyte nuclear factor 4 isoforms are encoded by distinct and differentially expressed genes. Mol Cell Biol **16**:925-931.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.

Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat Genet **24**:363-367.

Galili, U., and K. Swanson. 1991. Gene sequences suggest inactivation of alpha-1,3-galactosyltransferase in catarrhines after the divergence of apes from monkeys. Proc Natl Acad Sci U S A **88**:7401-7404.

Gibbs, R. A.G. M. WeinstockM. L. MetzkerD. M. MuznyE. J. SodergrenS. SchererG. ScottD. SteffenK. C. WorleyP. E. BurchG. OkwuonuS. HinesL. LewisC. DeRamoO. DelgadoS. Dugan-RochaG. MinerM. MorganA. HawesR. GillCeleraR. A. HoltM. D. AdamsP. G. AmanatidesH. Baden-TillsonM. BarnsteadS. ChinC. A. EvansS. FerrieraC. FoslerA. GlodekZ. GuD. JenningsC. L. KraftT. NguyenC. M. PfannkochC. SitterG. G. SuttonJ. C. VenterT. WoodageD. SmithH. M. LeeE. GustafsonP. CahillA. KanaL. Doucette-StammK. WeinstockK. FechtelR. B. WeissD. M. DunnE. D. GreenR. W. BlakesleyG. G. BouffardP. J. De JongK. OsoegawaB. ZhuM. MarraJ. ScheinI. BosdetC. FjellS. JonesM. KrzywinskiC. MathewsonA. SiddiquiN. WyeJ. McPhersonS. ZhaoC. M. FraserJ. ShettyS. ShatsmanK. GeerY. ChenS. AbramzonW. C. NiermanP. H. HavlakR. ChenK. J. DurbinA. EganY. RenX. Z. SongB. LiY. LiuX. QinS. CawleyA. J. CooneyL. M. D'SouzaK. MartinJ. Q. WuM. L. Gonzalez-GarayA. R. JacksonK. J. KalafusM. P. McLeodA. MilosavljevicD. VirkA. VolkovD. A. WheelerZ. ZhangJ. A. BaileyE. E. EichlerE. TuzunE. BirneyE. MonginA. Ureta-VidalC. WoodwarkE. ZdobnovP. BorkM. SuyamaD. TorrentsM. AlexanderssonB. J. TraskJ. M. YoungH. HuangH.

WangH. XingS. DanielsD. GietzenJ. SchmidtK. StevensU. VittJ. WingroveF. CamaraM. Mar AlbaJ. F. AbrilR. GuigoA. SmitI. DubchakE. M. RubinO. CouronneA. PoliakovN. HubnerD. GantenC. GoeseleO. HummelT. KreitlerY. A. LeeJ. MontiH. SchulzH. ZimdahlH. HimmelbauerH. LehrachH. J. JacobS. BrombergJ. Gullings-HandleyM. I. Jensen-SeamanA. E. KwitekJ. LazarD. PaskoP. J. TonellatoS. TwiggerC. P. PontingJ. M. DuarteS. RiceL. GoodstadtS. A. BeatsonR. D. EmesE. E. WinterC. WebberP. BrandtG. NyakaturaM. AdetobiF. ChiaromonteL. ElnitskiP. EswaraR. C. HardisonM. HouD. KolbeK. MakovaW. MillerA. NekrutenkoC. RiemerS. SchwartzJ. TaylorS. YangY. ZhangK. LindpaintnerT. D. AndrewsM. CaccamoM. ClampL. ClarkeV. CurwenR. DurbinE. EyrasS. M. SearleG. M. CooperS. BatzoglouM. BrudnoA. SidowE. A. StoneB. A. PayseurG. BourqueC. Lopez-OtinX. S. PuenteK. ChakrabartiS. ChatterjiC. DeweyL. PachterN. BrayV. B. YapA. CaspiG. TeslerP. A. PevznerD. HausslerK. M. RoskinR. BaertschH. ClawsonT. S. FureyA. S. HinrichsD. KarolchikW. J. KentK. R. RosenbloomH. TrumbowerM. WeirauchD. N. CooperP. D. StensonB. MaM. BrentM. ArumugamD. ShteynbergR. R. CopleyM. S. TaylorH. RiethmanU. MudunuriJ. PetersonM. GuyerA. FelsenfeldS. OldS. Mockrin, and F. Collins. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Giguere, V., N. Yang, P. Segui, and R. M. Evans. 1988. Identification of a new class of steroid hormone receptors. Nature **331**:91-94.

Glusman, G., I. Yanai, I. Rubin, and D. Lancet. 2001. The complete human olfactory subgenome. Genome Res **11**:685-702.

Goodwin, B., S. A. Jones, R. R. Price, M. A. Watson, D. D. McKee, L. B. Moore, C. Galardi, J. G. Wilson, M. C. Lewis, M. E. Roth, P. R. Maloney, T. M. Willson, and S. A. Kliewer. 2000. A regulatory cascade of the nuclear receptors FXR, SHP-1, and LRH-1 represses bile acid biosynthesis. Mol Cell **6**:517-526.

Harrison, P. M., D. Zheng, Z. Zhang, N. Carriero, and M. Gerstein. 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, S. B. 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., M. Girondot, P. Leroy, M. Clergue, H. Quach, M. Fellous, and T. Bourgeron. 2001. Transduction of the human gene FAM8A1 by endogenous retrovirus during primate evolution. Genomics **78**:38-45.

Johnson, M. E., L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E. E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature **413**:514-519.

King, D. C., J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison. 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., M. Nishikimi, T. Ozawa, and K. Yagi. 1988. Isolation and sequence analysis of a complementary DNA encoding rat liver L-gulono-gamma-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, W. H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Li, W. H., T. Gojobori, and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237-239.

Lindblad-Toh, K.C. M. WadeT. S. MikkelsenE. K. KarlssonD. B. JaffeM. KamalM. ClampJ. L. ChangE. J. Kulbokas, 3rdM. C. ZodyE. MauceliX. XieM. BreenR. K. WayneE. A. OstranderC. P. PontingF. GalibertD. R. SmithP. J. DeJongE. KirknessP. AlvarezT. BiagiW. BrockmanJ. ButlerC. W. ChinA. CookJ. CuffM. J. DalyD. DeCaprioS. GnerreM. GrabherrM. KellisM. KleberC. BardelebenL. GoodstadtA. HegerC. HitteL. KimK. P. KoepfliH. G. ParkerJ. P. PollingerS. M. SearleN. B. SutterR. ThomasC. WebberJ. BaldwinA. AbebeA. AbouelleilL. AftuckM. Ait-ZahraT. AldredgeN. AllenP. AnS. AndersonC. AntoineH. ArachchiA. AslamL. AyotteP. BachantsangA. BarryT. BayulM. BenamaraA. BerlinD. BessetteB. BlitshteynT. BloomJ. BlyeL. BoguslavskiyC. BonnetB. BoukhgalterA. BrownP. CahillN. CalixteJ. CamarataY. CheshatsangJ. ChuM. CitroenA. CollymoreP. CookeT. DawoeR. DazaK. DecktorS. DeGrayN. DhargayK. DooleyP. DorjeK. DorjeeL. DorrisN. DuffeyA. DupesO. EgbiremolenR. ElongJ. FalkA. FarinaS. FaroD. FergusonP. FerreiraS. FisherM. FitzGeraldK. FoleyC. FoleyA. FrankeD. FriedrichD. GageM. GarberG. GearinG. GiannoukosT. GoodeA. GoyetteJ. GrahamE. GrandboisK. GyaltsenN. HafezD. HagopianB. HagosJ. HallC. HealyR. HegartyT. HonanA. HornN. HoudeL. HughesL. HunnicuttM. HusbyB. JesterC. JonesA. KamatB. KangaC. KellsD. KhazanovichA. C. KieuP. KisnerM. KumarK. LanceT. LandersM. LaraW. LeeJ. P. LegerN. LennonL. LeuperS. LeVineJ. LiuX. LiuY. LokyitsangT. LokyitsangA. LuiJ. MacdonaldJ. MajorR. MarabellaK. MaruC. MatthewsS. McDonoughT. MehtaJ. MeldrimA. MelnikovL. MeneusA. MihalevT. MihovaK. MillerR. MittelmanV. MlengaL. MulrainG. MunsonA. NavidiJ. NaylorT. NguyenN. NguyenC. NguyenR. NicolN. NorbuC. NorbuN. NovodT. NyimaP. OlandtB. O'NeillK. O'NeillS. OsmanL. OyonoC. PattiD. PerrinP. PhunkhangF. PierreM. PriestA. RachupkaS. RaghuramanR. RameauV. RayC. RaymondF. RegeC. RiseJ. RogersP. RogovJ. SahalieS. SettipalliT. SharpeT. SheaM. SheehanN. SherpaJ. ShiD. ShihJ. SloanC. SmithT. SparrowJ. StalkerN. Stange-ThomannS. StavropoulosC. StoneS. StoneS. SykesP. TchuingaP. TenzingS. TesfayeD. ThoulutsangY. ThoulutsangK. TophamI. ToppingT. TsamlaH. VassilievV. VenkataramanA. VoT. WangchukT. WangdiM. WeiandJ. WilkinsonA. WilsonS. YadavS. YangX. YangG. YoungQ. YuJ. ZainounL. ZembekA. Zimmer, and E. S. Lander. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**:803-819.

Lower, R., J. Lower, and R. Kurth. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A **93**:5177-5184.

Lu, T. T., M. Makishima, J. J. Repa, K. Schoonjans, T. A. Kerr, J. Auwerx, and D. J. Mangelsdorf. 2000. Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors. Mol Cell **6**:507-515.

Maglich, J. M., A. Sluder, X. Guan, Y. Shi, D. D. McKee, K. Carrick, K. Kamdar, T. M. Willson, and J. T. Moore. 2001. Comparison of complete nuclear receptor sets from the human, Caenorhabditis elegans and Drosophila genomes. Genome Biol **2**:RESEARCH0029.

Maston, G. A., and M. Ruvolo. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol **19**:320-335.

Ohshima, K., M. Hattori, T. Yada, T. Gojobori, Y. Sakaki, and N. Okada. 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., H. Kranz, I. Kober, P. Thompson, M. Hoefer, B. Haubold, B. Remmel, H. Voss, C. Kaiser, M. Albers, Z. Cheruvallath, D. Jackson, G. Casari, M. Koegl, S. Paabo, J. Mous, C. Kremoser, and U. Deuschle. 2003. Identification of farnesoid X receptor beta as a novel mammalian nuclear receptor sensing lanosterol. Mol Cell Biol **23**:864-872.

Pare, J. F., D. Malenfant, C. Courtemanche, M. Jacob-Wagner, S. Roy, D. Allard, and L. Belanger. 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., A. S. Carpentier, M. Duffraisse, and V. Laudet. 2001. How many nuclear hormone receptors are there in the human genome? Trends Genet **17**:554-556.

Roy, S. W., and W. Gilbert. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet **7**:211-221.

Shemesh, R., A. Novik, S. Edelheit, and R. Sorek. 2006. Genomic fossils as a snapshot of the human transcriptome. Proc Natl Acad Sci U S A **103**:1364-1369.

Sladek, R., B. Beatty, J. Squire, N. G. Copeland, D. J. Gilbert, N. A. Jenkins, and V. Giguere. 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, A. E., S. W. Mathews, D. Hough, V. P. Yin, and C. V. Maina. 1999. The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. Genome Res **9**:103-120.

Stedman, H. H., B. W. Kozyak, A. Nelson, D. M. Thesier, L. T. Su, D. W. Low, C. R. Bridges, J. B. Shrager, N. Minugh-Purvis, and M. A. Mitchell. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature **428**:415-418.

Taraviras, S., T. Mantamadiotis, T. Dong-Si, A. Mincheva, P. Lichter, T. Drewes, G. U. Ryffel, A. P. Monaghan, and G. Schutz. 2000. Primary structure, chromosomal mapping, expression and transcriptional activity of murine hepatocyte nuclear factor 4gamma. Biochim Biophys Acta **1490**:21-32.

Tchenio, T., E. Segal-Bendirdjian, and T. Heidmann. 1993. Generation of processed pseudogenes in murine cells. Embo J **12**:1487-1497.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. Genome Res **13**:2559-2567.

Wang, W., F. G. Brunet, E. Nevo, and M. Long. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci U S A **99**:4448-4453.

Waterston, R. H.K. Lindblad-TohE. BirneyJ. RogersJ. F. AbrilP. AgarwalR. AgarwalaR. AinscoughM. AlexanderssonP. AnS. E. AntonarakisJ. AttwoodR. BaertschJ. BaileyK. BarlowS. BeckE. BerryB. BirrenT. BloomP. BorkM. BotcherbyN. BrayM. R. BrentD. G. BrownS. D. BrownC. BultJ. BurtonJ. ButlerR. D. CampbellP. CarninciS. CawleyF. ChiaromonteA. T. ChinwallaD. M. ChurchM. ClampC. CleeF. S. CollinsL. L. CookR. R. CopleyA. CoulsonO. CouronneJ. CuffV. CurwenT. CuttsM. DalyR. DavidJ. DaviesK. D. DelehauntyJ. DeriE. T. DermitzakisC. DeweyN. J. DickensM. DiekhansS. DodgeI. DubchakD. M. DunnS. R. EddyL. ElnitskiR. D. EmesP. EswaraE. EyrasA. FelsenfeldG. A. FewellP. FlicekK. FoleyW. N. FrankelL. A. FultonR. S. FultonT. S. FureyD. GageR. A. GibbsG. GlusmanS. GnerreN. GoldmanL. GoodstadtD. GrafhamT. A. GravesE. D. GreenS. GregoryR. GuigoM. GuyerR. C. HardisonD. HausslerY. HayashizakiL. W. HillierA. HinrichsW. HlavinaT. HolzerF. HsuA. HuaT. HubbardA. HuntI. JacksonD. B. JaffeL. S. JohnsonM. JonesT. A. JonesA. JoyM. KamalE. K. KarlssonD. KarolchikA. KasprzykJ. KawaiE. KeiblerC. KellsW. J. KentA. KirbyD. L. KolbeI. KorfR. S. KucherlapatiE. J. KulbokasD. KulpT. LandersJ. P. LegerS. LeonardI. LetunicR. LevineJ. LiM. LiC. LloydS. LucasB. MaD. R. MaglottE. R. MardisL. MatthewsE. MauceliJ. H. MayerM. McCarthyW. R. McCombieS. McLarenK. McLayJ. D. McPhersonJ. MeldrimB. MeredithJ. P. MesirovW. MillerT. L. MinerE. MonginK. T. MontgomeryM. MorganR. MottJ. C. MullikinD. M. MuznyW. E. NashJ. O. NelsonM. N. NhanR. NicolZ. NingC. NusbaumM. J. O'ConnorY. OkazakiK. OliverE. Overton-LartyL. PachterG. ParraK. H. PepinJ. PetersonP. PevznerR. PlumbC. S. PohlA. PoliakovT. C. PonceC. P. PontingS. PotterM. QuailA. ReymondB. A. RoeK. M. RoskinE. M. RubinA. G. RustR. SantosV. SapojnikovB. SchultzJ. SchultzM. S. SchwartzS. SchwartzC. ScottS. SeamanS. SearleT. SharpeA. SheridanR. ShownkeenS. SimsJ. B. SingerG. SlaterA. SmitD. R. SmithB. SpencerA. StabenauN. Stange-ThomannC. SugnetM. SuyamaG. TeslerJ. ThompsonD. TorrentsE. TrevaskisJ. TrompC. UclaA. Ureta-VidalJ. P. VinsonA. C. Von NiederhausernC. M. WadeM. WallR. J. WeberR. B. WeissM. C. WendlA. P. WestK. WetterstrandR. WheelerS. WhelanJ. WierzbowskiD. WilleyS. WilliamsR. K. WilsonE. WinterK. C. WorleyD. WymanS. YangS. P. YangE. M. ZdobnovM. C. Zody, and E. S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Zhang, Z., P. E. Burch, A. J. Cooney, R. B. Lanz, F. A. Pereira, J. Wu, R. A. Gibbs, G. Weinstock, and D. A. Wheeler. 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., and M. Gerstein. 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., P. Harrison, and M. Gerstein. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res **12**:1466-1482.

Zhang, Z., P. M. Harrison, Y. Liu, and M. Gerstein. 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., A. Frankish, R. Baertsch, P. Kapranov, A. Reymond, S. W. Choo, Y. Lu, F. Denoeud, S. E. Antonarakis, M. Snyder, Y. Ruan, C.-L. Wei, T. R. Gingeras, R. Guigo, J. Harrow, and M. B. Gerstein. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription and evolution. Genome Res:In press.

Zheng, D., Z. Zhang, P. M. Harrison, J. Karro, N. Carriero, and M. Gerstein. 2005. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol **349**:27-45.

Tables

Table 1. Symbols of NR used in the text

Symbol	Official name	Full name
FXRβ	NR1H5	Farnesoid X receptor, beta
HNF4γ	NR2A2	Hepatocyte nuclear factor 4, gamma
ERRα	NR3B1	Estrogen-related receptor, alpha
Rev-erbβ	NR1D2	Thyroid hormone receptor, alpha-like
PNR	NR2E3	Photoreceptor-specific nuclear receptor
ERRβ	NR3B2	Estrogen-related receptor, beta
LRH1	NR5A2	Liver receptor homolog 1

Table 2. Human and rodent nuclear receptor pseudogenes

Genome	Pseudogene	Accession [1]	Location [2]		Type	Truncation [3]	
			Chr. band	Coordinate		5'	3'
Human	*ψFXRβ*	15259	1p13.1+	115181466	unitary	no	no
	ψHNF4γ	128390	13q21.1−	55471366	processed	yes	yes
	ψERRa	5316	13q12.11−	19032156	processed	no	no
	ψERRa	24162	13q12.11+	20732460	processed	no	no
Chimp	*ψFXRβ*	8400	1−	122802667	unitary	no	no
	ψHNF4γ	8401	13−	55892954	processed	yes	yes
	ψERRa	8402	13−	19079069	processed	no	no
Mouse	*ψRev-erbβ*	19393	19qC3+	40244011	semiprocessed	no	no
	ψPNR	6324	15qB3.1+	35678192	processed	yes	no
	ψERRβ	10804	XqA5+	57351250	processed	no	no
	ψLRH1	8260	3qH2+	144716412	semiprocessed	yes	no
	ψLRH1	17110	6qF1−	118583245	processed	yes	no
Rat	*ψERRβ*	8720	Xq36+	146717523	processed	no	no
	ψLRH1	1916	11q21+	48578386	processed	yes	no
	ψLRH1	17561	Xq14−	30976310	processed	yes	no

[1] 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.

[2] 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.

[3] Truncation is relative to the coding sequences. 5' and 3' refer to the ends of the coding sequence of the functional parent gene.

Figure legends

Figure 1. The gene structure of human *ψFXRβ*. The mouse FXRβ protein sequence [9] and the translation of the human genomic sequence at the *ψFXRβ* 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 *ψFXRβ* 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).

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

Figure 3. Human, chimpanzee, and rhesus *ψFXRβ*. (A) Disruptive defects in *ψFXRβ*. Such sequence defects, including frame shifts, nonsense mutations, and splice site mutations, were found in the sequence alignment at 14 orthologous positions, which are numbered and accented in black bold underlined letters. For clarity, the base letters in chimpanzee and rhesus *ψFXRβ* sequences identical to their corresponding ones in human *ψFXRβ* 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 *ψFXRβ*. 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 descendents. Thus defects 1, 2, and 14 are found in all three primate species, while defects 3, 4, 5, 9, and 10 are found in both human and chimpanzee but not in rhesus.

Figure 4. The evolution of *FXR* and *FXRβ*. (A) The relationships and divergence times of major groups of vertebrates.(Hedges 2002) Both the *FXR-FXRβ* duplication and *FXRβ* inactivation events are dated and marked accordingly in the phylogeny. Branch lengths are not proportional to time. (B) Dendrogram of *FXR* and *FXRβ*. The evolution of *FXR* and *FXRβ* in mammals is juxtaposed and highlighted in the tree. The difference in their evolution speed is readily perceivable. Branch lengths are proportional to time. The dendrogram was tested with a bootstrap of 1000 replications and the bootstrap values in percentage are labeled by the branching points.

Figure 5. Conservation of intergenic sequence upstream to human *ψFXRβ*. (A) Three highly conserved sequence segments immediately upstream to the 'coding sequence' of *ψFXRβ* and the alignment of orthologous sequences from 13 vertebrates in these three sequence segments. (B) A highly conserved ~250 bp sequence segment with a high regulatory potential 4.5 Kb upstream to *ψFXRβ* and a zoom-in view of (C). Notice that this sequence segment has a high regulatory potential comparable to that of the transcription start site of the functional gene *SYCP1*.

Figure 6. Detailed structures of two 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 two short interspersed elements (SINEs; the white arrows) and one long interspersed element (LINE; the gray arrow) were found. These three interspersed repetitive sequences were not found in the intron at the same location in the functional paralogous gene. The similar sequences shared between the two introns, enlarged for clarity, are indicated by thicker line segments. In the picture only the exons and the features in the two introns of interest were kept in proportion within each group. (B) The remnant intron in mouse *ψLRH1* on chromosome 3. Sequence alignment shows that two 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.

Figure 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.

Figures

```
                                   ▽      ▽
Mouse FXRβ      1 MANTYVATSDGYYLAEP-TQYYD
                  ||||||   | | ||||    ||
                  MANTYVTACDRYCLAEP!VHCYD ^
Human ψFXRβ  3327 agaatgagtgattcggcggcttg[ATA...Intron 1...TAG]at
                  tcacatccgagagtcacttaga
                  gatttctattgttttaacgtct
                                   a

Legend
  ▽    Nonfunctionalization site
  !    Frame shift
  *    Nonsense mutation
  ^    Splice site mutation
  [ ]  Intron boundaries

                                                       ▽
Mouse FXRβ     23 ILPEQFHYQLCDTDFQEPPYCQYSTAQFPPALQSPSLQSHFNTHGLDPQYSGGSWCGLDARESGQSTYVVVHDDEDEFPGAQRCRAT-CSLRWKGQD
                  |||||: ||| || |:: |||||| ||||||||| || :||||: |||| | |  ||  :||  : |||| || || :|| :| | |  | | |||:
                  ILPEQISYQLHDTHFKKSPYCQYSIAQFPPALQSESL*NHFNTYRLDPQDSDGGQCGFSSRELNKPTYVVAHDAEDGYPGIKRSRPTYSSSRNKGQE
Human ψFXRβ  6184 atcgcaatcccgactaatcttcttagctccgtctgtttactaatatgccgagggctgtatcgtaacatgggcgggggtcgaaatacattttaaagcg
                  ttcaatgaataacataaccagaactcatccctacactaaatacagtacaagaggaggtgcgataaccattcaacaagacgtagcgcccacccgaagaa
                  taaaatttggtctttaaattcgttttgttatagtataattccttagtagctttagtattttaatacttgtttttatactaaagcaacttcgatgaga

                                                                 ▽           ▽
Mouse FXRβ    119 DMLCMVCGDKA--SGYHYNALTCEGCKG                        FFRRSITKNAVYSCKNGGHCEMDMYMRRKCQ
                  :: |:||||||  | |||||||||||||                        ||: ||  ||||||:|| |||||||||||||
                  EF-CVVCGDKASPSPYHYNALTCEGCKG                   ^    FFQCSIN!NAVYSCRNGSHCEMDMYMRRKCQ
Human ψFXRβ  6475 gt tggtggagtctctctagcatggtag[GTA...Intron 2...TGG]gtttctaaaaaggtataagactgagatacaatc
                  at gttggaacccccaaaactcgagga                         ttaggtaaactagggaggagatatatggaga
                  ac tattttaaaaaatttatctatca                          ttatccc taattcgtttctagcgcgtaata

Mouse FXRβ    176 ECRLKKCKAVGMLAEC                        LLTEIQCKSKRLRKNFKHGPALYPAIQVEDEGADTKHVSSSTRSGKG
                  |||||| |||||||||                        |||||||| |||:||||     |  |:||:|| |   :||:|| ||
                  ECRLKKYKAVGMLAEC                        LLTEIQCKLKRLQKNFKEKNHFYSNIKVEEEGVDHSFLSSTTRPGK-
Human ψFXRβ  8222 gtacaatagggatggt[GTA...Intron 3...CAG]gttcagactataaccaatagaactttaaagggggggcatcttaaacga
                  aggtaaaactgttca                         ttcatagatagtaaataaaaatacatataaagtaagttccccgcga
                  gcagagtgaaaggaa                         gcaacataagatagctgggtttctccagagaaaccttaacctataa

Mouse FXRβ    239                       VQDNMTLTQEEHRLLNTIVTAHQKSMIPLGETSKL                      LQEGSNPELSFLRLSEV
                                        :|::| ||:|||:|:| || ||||  ||| ||:                        ||| :|||||||:|||
                                        IQESMELTEEEHQLINNIVAAHQKYTIPLEETNLY                      LQEHTNPELSFLQLSET
Human ψFXRβ 14749 [GTG...Intron 4...TAG]acgaagcagggcccaaaagggccataactggaatt[GTG...Intron 5...TAG]ccgcaacgcattcctga
                                        taagtatcaaaaattaattccaaaactctaacata                      taaacacatgttatcac
                                        tgacgaatagatgcttctgtttaatcttaaaatgt                      ggatattagctgacaga

Mouse FXRβ    291 SVLHIQGLMKFTKGLPG                        FENLTTEDQAALQKASKTEVMFLHVAQLYGGKDSTSG
                  :||||:||| |||||||                        ||||  ||| |||| |||||:||| ||||  | | |
                  AVLHIRGLMNFTKGLPG                        FENLANEDQTALQKGSKTEVIFLHGAQLYSQKQSASE
Human ψFXRβ 15136 ggccacgcaataagccg[GTA...Intron 6...AAG]gatgatgaggcagccagtaaggatccggcctacactgtg
                  cttatggttatcagtc                         taatcaaaacctaagcacattttagcatagaaaccca
                  acacatgagttcggca                         tatgctgtataaggaaatagatctgcatctgaaacta

                                                                                                        ▽
Mouse FXRβ    345 S                        TMRPAKPSAGTLEVHNPSADES-VHSPENFLKEGYPSAPLTD                        IT
                  |                        ::|    |  |   || | | | : | | |  |  ||  |                          :|
                  S                        SVRILNHSDYTPNCHNRSGDRSLICSMEKFYNEECPSTTLIG                   ^    VT
Human ψFXRβ 16863 a[GTA...Intron 7...TAG]gttgaatactgtacatcaaaggaacattagattaggtctaacag[GTA...Intron 8...TGG]gtga
                                         ctgttaacaaccagaagggaggttgctaataaaagcccctt                         tc
                                         tgaaattattaattctgtttatttttgaatctaatttttat                         tt

                       ▽                                                                   ▽
Mouse FXRβ    389 KEFIASLSYFYRRMSELHVSDTEYALLTATTVLFSD                        RPCLKNKQHIENLQEPVLQLLFK
                   |||  | |||:|||:| |::|||||| | |::|||                        ||||||||::||| |||||:|:|
                  EEFIT!LFYFYKRMSKLDVTNTEYALL-AATIVFSD                        RPCLKNKQYMENL*EPVLQILYK
Human ψFXRβ 17540 ggtaa cttttaaaaacggaaagtgcc ggaagttg[GTA...Intron 9...CAG]atcctcaaactagattgcgtcatta
                  aattccttataagtgatatcacaactt ccctttc                         gcgtaaaaataataacttattaa
                  aattaagtcccaagcattatttattgt aaattta                         tactatgatgataaaataaagtg

                            ▽                                        ▽
Mouse FXRβ    448 FSKMYHPEDPQHFAHLIGRLTELRTLSHSHSEILRMWKTKDPRLVMLFSEKWDLHSFS 505
                  :||||||||| |||||| : ||||||:::|||||  ||||||:|  |:|||  |:| :
                  YSKMYHPEDP*HFAHLIWKHTELRTLNYNHSEILSTWKTKDPKLATLLSEK*NLYS-C
Human ψFXRβ 20183 ttaatccggctctgccatacagcaacatactgacaataaagcatgatctgatattt t
                  acataacaacaatcattgaacatgctaaaacattgcgacaacatccttcaagatac g
                  taagttaacagttctcaggttagatgttctaaatctgaagccagttactggatgtt c 20352
```

Figure 1.

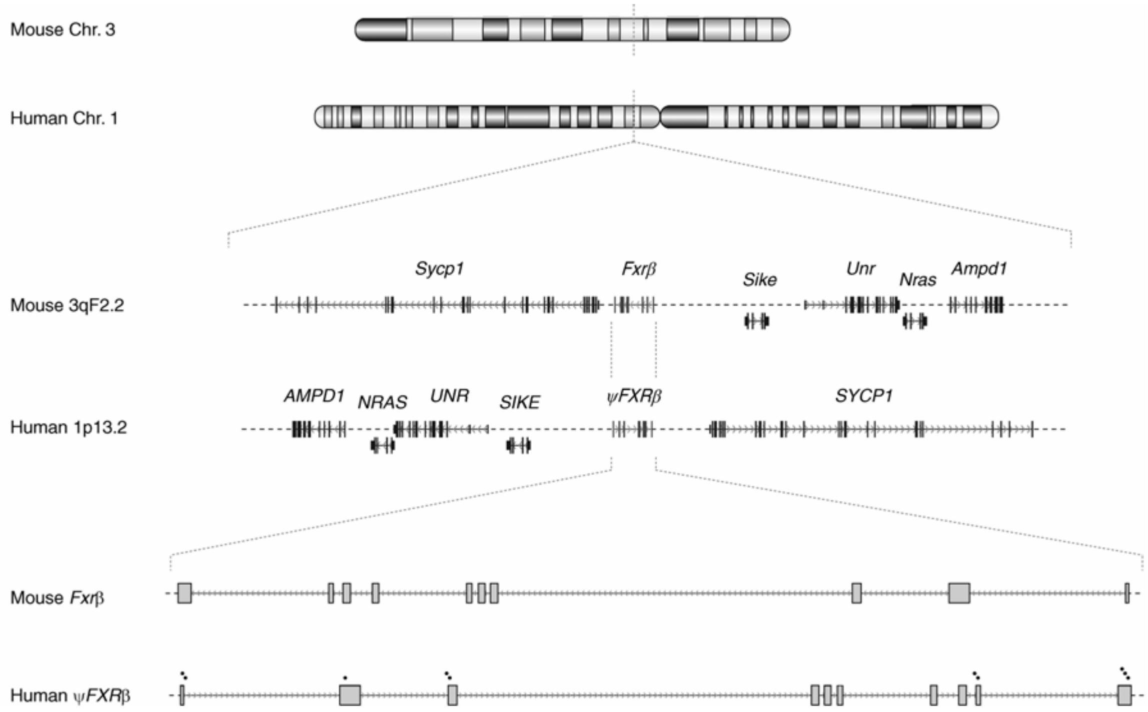

Mouse Chr. 3
Human Chr. 1
Sycp1
Fxrβ
Sike
Unr
Nras
Ampd1
Mouse 3qF2.2
AMPD1
NRAS
UNR
SIKE
ψFXRβ
SYCP1
Human 1p13.2
Mouse Fxrβ
Human ψFXRβ

Figure 2.

```
                                                                       (1)              (2)
Human  ψFXRβ        atggcaaatacttatgtcactgcatgtgataggtattgtcttgctgaaccagtcagtgcattgctatg[ATA...Intro
Chimp  ψFXRβ        ...................................................gtca.............[A..........
Rhesus ψFXRβ        ...............------....c.........................gtca.............[A..........

n1...TAG]atattttaccagaacaaattagttatcagctgcatgacactcattttaaaaaatcaccttattgccagtattctattgctcagtttcctcc
........]...........................................................................................
.....~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

                      (3)
agctttacagtctgaatctttataaaatcatttcaacacttatagattggatccacaggacagtgatggtggacagtgtggatttagttctcgtgaatta
......................taa.g.........................................................................
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

aataaacccacttatgtggttgctcatgatgctgaagatggataccctggaataaaaaggtccagaccaacctattcttcctcgagaaataagggacagg
....................................................................................................
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                                                                  (4)
aagaattctgtgtagtttgtggtgataaagcatcaccatcaccatatcattataatgcacttacctgtgaaggttgcaaag[GTA...Intron2...TG
...........................................g.....................................[.................G
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                        (5)
G]gtttttttcaatgtagcatcaacaaaatgcagtatatagttgcaggaatggtagtcactgtgaaatggacatgtacatgcgtagaaaatgtcaagagt
.]...c...................aa.....g...................................................................
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                           (6)
gcagactgaaaaagtataaggcagtaggaatgttggcagaat[GTA...Intron3...CAG]gtttgctcacagaaatccaatgtaaattaaagagact
...............g..........................[...................].....................................
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~.G.]...........................cg.g......
                         (7)
tcaaaagaactttaaggagaagaatcatttttactctaacatcaaagtggaagaggaaggagtagaccacagttttctatcatccaccactagacctgga
...................................................................a................................
................c....t...tga..............g..............g.........a...ca................g....t.....
   (8)
aaa[GTG...Intron4...TAG]attcaggaaagcatggaactaactgaagaggaacatcagctcattaataacattgtggctgctcatcaaaaatata
...[A..................]............................................................................
...[...................]........................c...................................................

ccattcctttagaagaaacaaatttgtat[GTG...Intron5...TAG]ctgcaggaacatacaaatcctgaactgagctttttgcaactctcagagac
.............................[...................]............g.....................................
...........................t.[...................]............g.....t......................t........

agcagtcctacacatacgtgggctaatgaattttaccaaggggctcccag[GTA...Intron6...AAG]gatttgaaaatttggccaatgaggatcaa
..................................................[...................].............................
................t.......................a.........[...................]..............a...g..........

actgcactacagaagggatcaaaaactgaagtgatatttctccatggggcccaactttacagtcagaaacaatcagcctctgaaa[GTA...Intron7.
.....................................................................................[..............
...................................g................c................g......a........[..............

..TAG]gttctgtgagaatattaaatcattcagattatacaccaaattgtcacaataggagtggtgatagaagtcttatttgttctatggaaaaatttta
.....].............................................................................................g
.....]a...................................tg..........g............c...........c................g...
                                                 (9)                    (10)
caatgaagaatgtccttctactactctaattg[GTA...Intron8...TGG]gtgttactgaagaatttattac-acactgttttacttctacaaaaga
................................[.................G.]....................ac.........................
.................a...........c..[.................A.]..a...................c...................g....

atgagcaaacttgatgtaactaatactgaatatgctctgcttgcagcaacaattgttttttcag[GTA...Intron9...CAG]atcgtccatgcctta
................................................................[...................]...............
..........c....act.....................a........................[...................]...............
          (11)           (12)                                                        (13)
aaaataagcaatatatggaaaatttataagaaccagttttacaaatattgtataagtattcaaaaatgtatcatccagaagacccatagcattttgccca
...........taa............c...........................................................c..........t..
.g........................g............................a.....................a........c...t......tt.

tctcatatggaagcatactgaactgagaactctgaattataaccattcagaaatacttagcacttggaaaacaaaggaccccaaattggctactttactc
....................................c...................................c...........................
.............tt...........a............c............................................................
        (14)
tctgagaagtgaaatttgtattcttgc
.........tga...............
.........tga......c.....c..
```

Figure 3A.

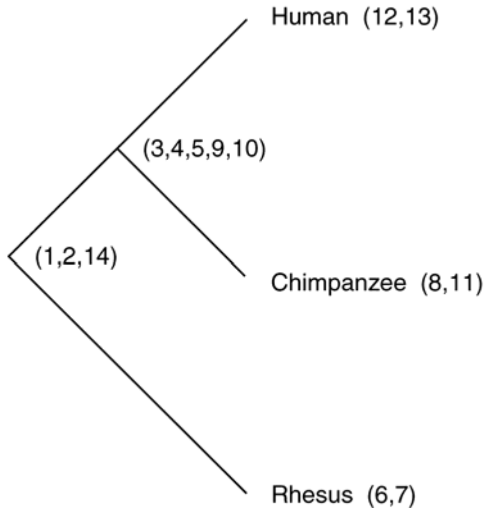


Figure 3B.

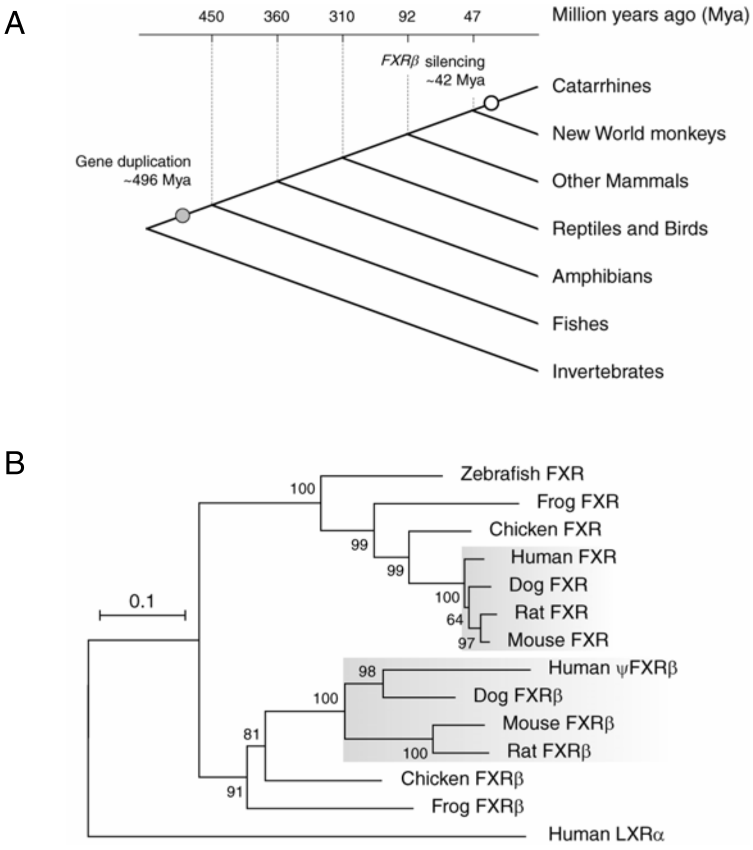

A
450
360
310
92
47
Million years ago (Mya)
FXRβ silencing ~42 Mya
Gene duplication ~496 Mya
Catarrhines
New World monkeys
Other Mammals
Reptiles and Birds
Amphibians
Fishes
Invertebrates
B
0.1
100
99
99
100
64
97
Zebrafish FXR
Frog FXR
Chicken FXR
Human FXR
Dog FXR
Rat FXR
Mouse FXR
98
100
81
100
91
Human ψFXRβ
Dog FXRβ
Mouse FXRβ
Rat FXRβ
Chicken FXRβ
Frog FXRβ
Human LXRα

Figure 4.

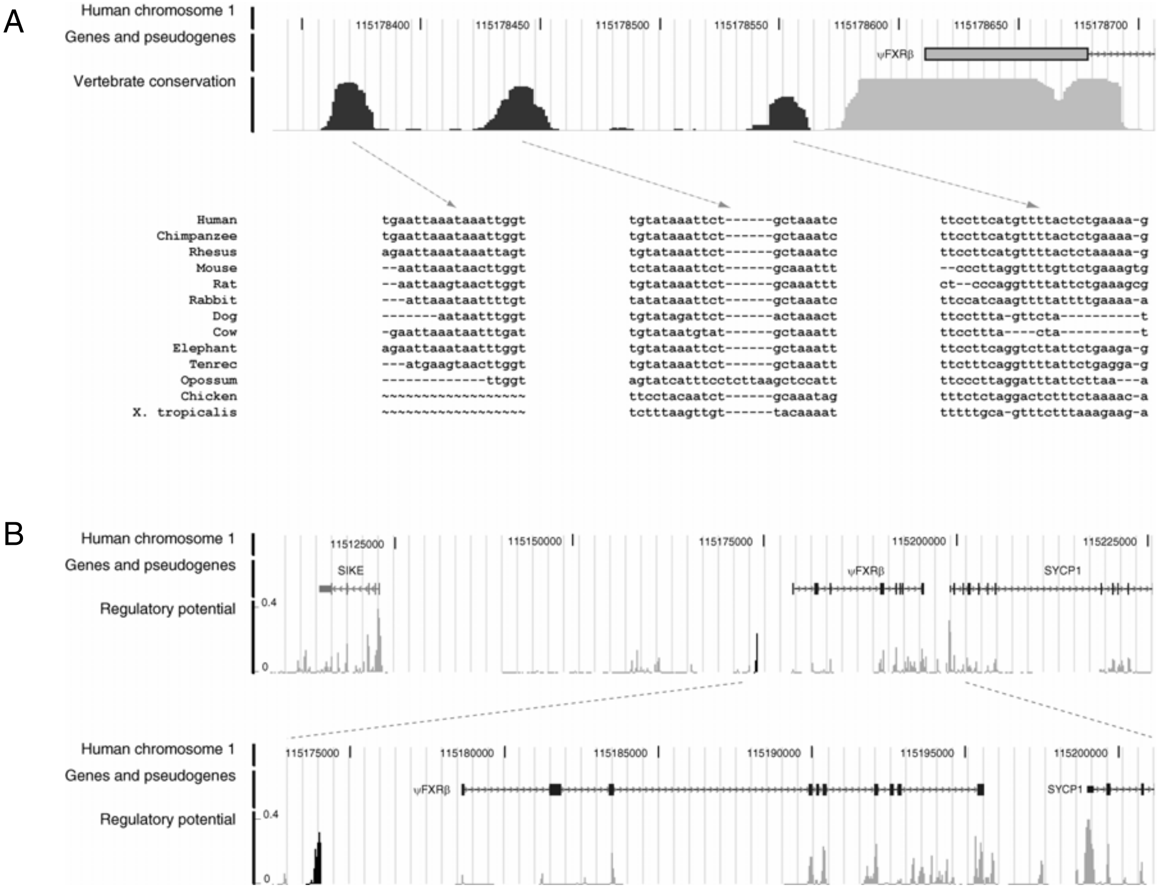

A
Human chromosome 1
Genes and pseudogenes
Vertebrate conservation
115178400
115178450
115178500
115178550
115178600
115178650
115178700
ψFXRβ
Human tgaattaaataaattggt tgtataaattct------gctaaatc ttccttcatgttttactctgaaaa-g
Chimpanzee tgaattaaataaattggt tgtataaattct------gctaaatc ttccttcatgttttactctgaaaa-g
Rhesus agaattaaataaattagt tgtataaattct------gctaaatc ttccttcatgttttactctaaaaa-g
Mouse --aattaaataacttggt tctataaattct------gcaaattt --cccttaggttttgttctgaaagtg
Rat --aattaagtaacttggt tgtataaattct------gcaaattt ct--cccaggttttattctgaaagcg
Rabbit ---attaaataattttgt tatataaattct------gctaaatc ttccatcaagttttattttgaaaa-a
Dog -------aataatttggt tgtatagattct------actaaact ttccttta-gttcta----------t
Cow -gaattaaataatttgat tgtataatgtat------gctaaatt ttccttta----cta----------t
Elephant agaattaaataatttggt tgtataaattct------gctaaatt ttccttcaggtcttattctgaaga-g
Tenrec ---atgaagtaacttggt tgtataaattct------gctaaatt ttctttcaggttttattctgagga-g
Opossum -------------ttggt agtatcatttcctcttaagctccatt ttcccttaggatttattcttaa---a
Chicken ~~~~~~~~~~~~~~~~~~ ttcctacaatct------gcaaatag tttctctaggactctttctaaaac-a
X. tropicalis ~~~~~~~~~~~~~~~~~~ tctttaagttgt------tacaaaat tttttgca-gtttctttaaagaag-a
B
Human chromosome 1
Genes and pseudogenes
Regulatory potential
115125000
115150000
115175000
115200000
115225000
SIKE
ψFXRβ
SYCP1
0.4
0
Human chromosome 1
Genes and pseudogenes
Regulatory potential
115175000
115180000
115185000
115190000
115195000
115200000
ψFXRβ
SYCP1
0.4
0

Figure 5.

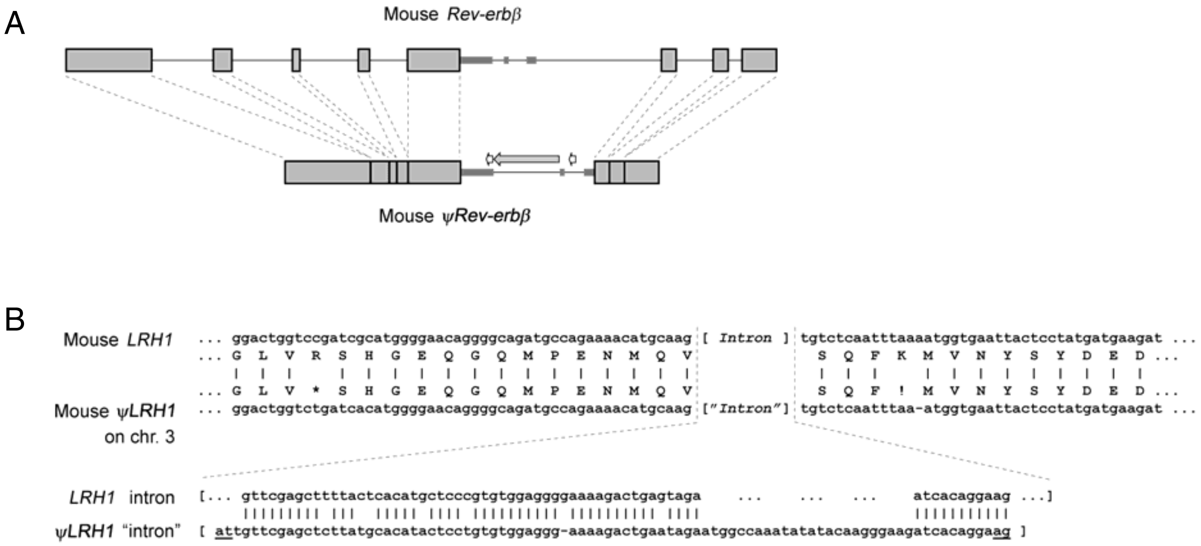

A
Mouse Rev-erbβ
Mouse ψRev-erbβ
B
Mouse LRH1
Mouse ψLRH1
on chr. 3
... ggactggtccgatcgcatggggaacaggggcagatgccagaaaacatgcaag [Intron] tgtctcaatttaaaatggtgaattactcctatgatgaagat ...
... G L V R S H G E Q G Q M P E N M Q V S Q F K M V N Y S Y D E D ...
... G L V * S H G E Q G Q M P E N M Q V S Q F ! M V N Y S Y D E D ...
... ggactggtctgatcacatggggaacaggggcagatgccagaaaacatgcaag ["Intron"] tgtctcaatttaa-atggtgaattactcctatgatgaagat ...
LRH1 intron
ψLRH1 "intron"
[... gttcgagcttttactcacatgctcccgtgtggaggggaaaagactgagtaga ... ... atcacaggaag ...]
[attgttcgagctcttatgcacatactcctgtgtggaggg-aaaagactgaatagaatggccaaatatatacaagggaagatcacaggaag]

Figure 6.

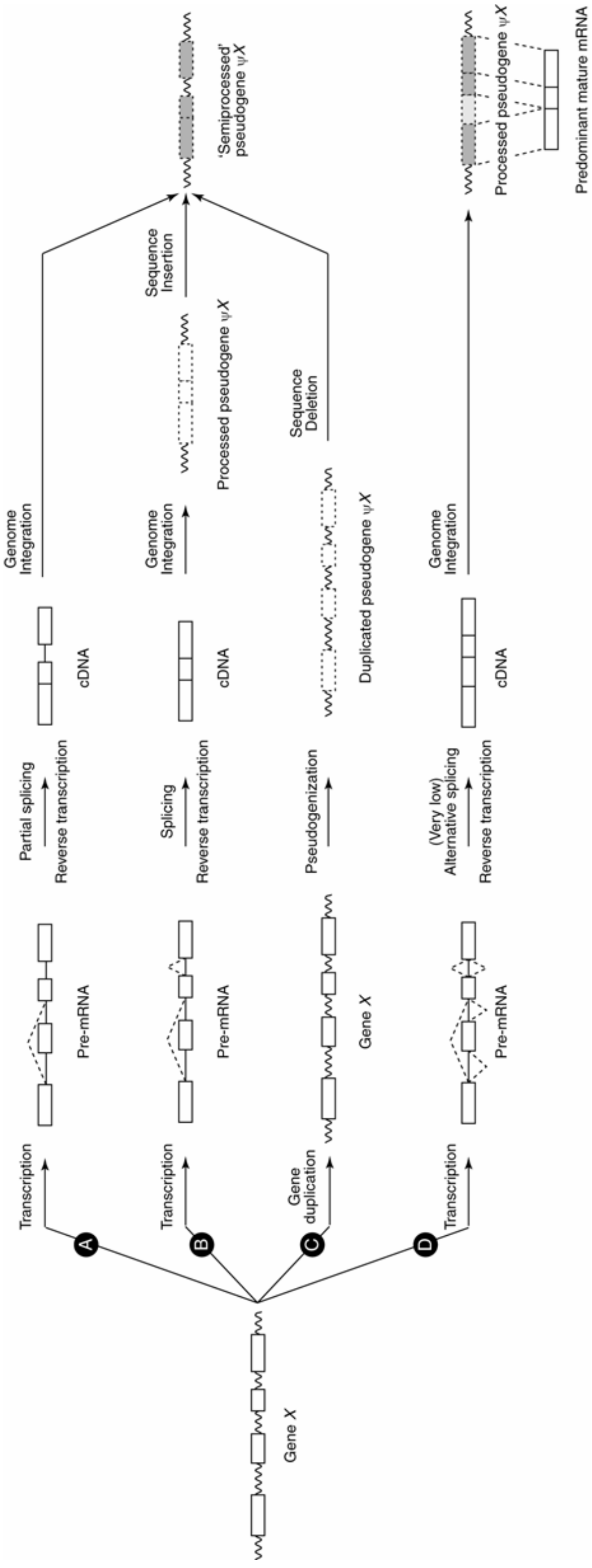

Gene X
A
Transcription
Pre-mRNA
Partial splicing
Reverse transcription
cDNA
Genome Integration
B
Transcription
Pre-mRNA
Splicing
Reverse transcription
cDNA
Genome Integration
Processed pseudogene ψX
Sequence Insertion
C
Gene duplication
Gene X
Pseudogenization
Duplicated pseudogene ψX
Sequence Deletion
'Semiprocessed' pseudogene ψX
D
Transcription
Pre-mRNA
(Very low) Alternative splicing
Reverse transcription
cDNA
Genome Integration
Processed pseudogene ψX
Predominant mature mRNA

Figure 7.

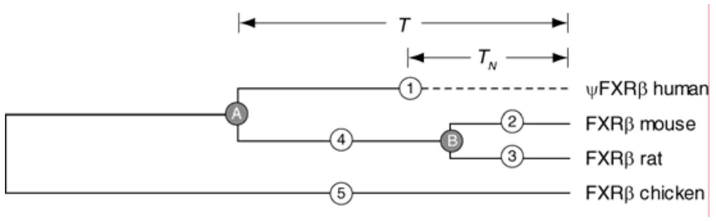
Supplementary materials

Supplementary table1. Dating the *FXR-FXRβ* duplication event

Ka (nonsynonymous substitution per nonsynonymous site)				
	FXR_{mouse}	FXR_{rat}	$FXR\beta_{mouse}$	$FXR\beta_{rat}$
FXR_{mouse}	—			
FXR_{rat}	0.015 ± 0.004	—		
$FXR\beta_{mouse}$	0.458 ± 0.035	0.468 ± 0.036	—	
$FXR\beta_{rat}$	0.466 ± 0.033	0.471± 0.035	0.062 ± 0.009	—
Ks (synonymous substitution per synonymous site)				
FXR_{mouse}	—			
FXR_{rat}	0.278 ± 0.035	—		
$FXR\beta_{mouse}$	7.835 ± 4.676	9.650 ± 5.331	—	
$FXR\beta_{rat}$	2.897 ± 0.736	5.885 ± 3.688	0.231 ± 0.030	—

1. The method used to date the gene duplication event (Li 1997) uses *FXR* and *FXRβ* sequences, each of which comes from two species respectively. Among the small number of available sequences, we chose to use those from mouse and rat for a sensible degree of sequence divergence and also for the good estimate of the species divergence time (T_S) between them.
2. Since the synonymous substitutions per synonymous site (*Ks*) are large and thus cannot be estimated accurately, they are not used to calculate T_D. As the equation shows below, only the nonsynonymous substitution per nonsynonymous site (*Ka*) are used.
3. The method assumes a constant substitution rate at least since the duplication event. To test the constant synonymous substitution rate condition on which the following calculation is based, we compared *Ka* of FXR_{mouse}-$FXR\beta_{mouse}$ and *Ka* of FXR_{rat}-$FXR\beta_{rat}$. The assumption of a constant rate seems reasonable, as the difference between them is small (|0.458−0.471| = 0.013).
4. $$T_D = 2 \cdot T_S \cdot \frac{\overline{K}_{FXR,FXR\beta}}{K_{FXR} + K_{FXR\beta}} = 2 \cdot 41 \cdot \frac{(0.458 + 0.468 + 0.466 + 0.471)/4}{0.015 + 0.062} = 496 \text{ (Mya)}$$

Supplementary table 2. Dating the *ψFXRβ* nonfunctionalization event



Lineage	*t*	*N*	*S*	*ω* =*Ka*/*Ks*	*Ka*	*Ks*	*N*×*Ka*	*S*×*Ks*
1	0.299	1151.6	438.4	0.6034	0.0843	0.1398	97.1	61.3
2	0.132	1151.6	438.4	0.3642	0.0298	0.0818	34.3	35.9
3	0.198	1151.6	438.4	0.2329	0.0345	0.1482	39.8	65
4	0.487	1151.6	438.4	0.3864	0.1129	0.2921	130	128.1
5	1.786	1151.6	438.4	0.1024	0.1742	1.7015	200.6	746

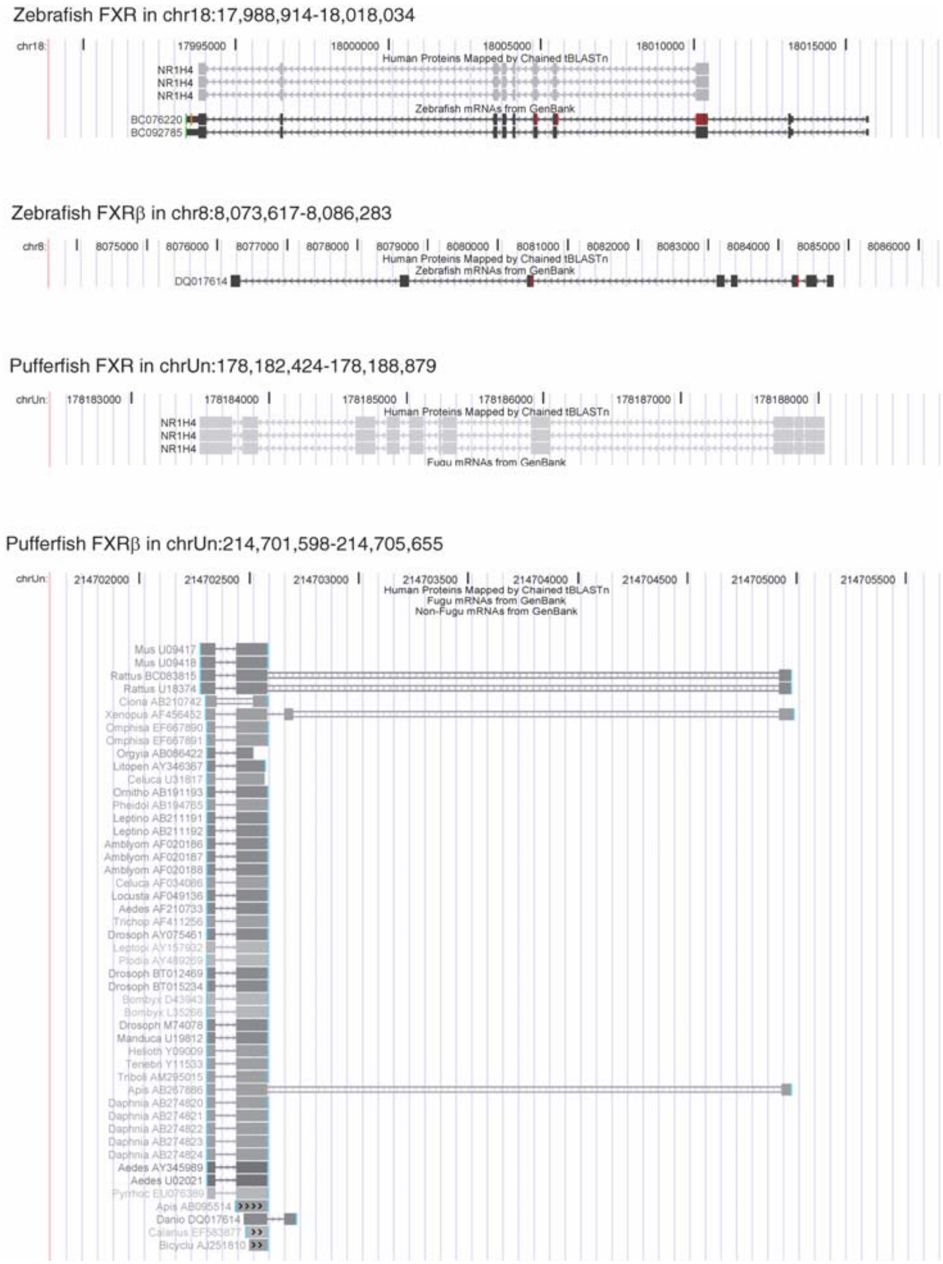
1. The method used to date the *ψFXRβ* nonfunctionalization event (Chou et al. 2002) assumes that non-synonymous mutations are selected against until the gene is inactivated; thereafter mutations at both synonymous and non-synonymous sites accumulate at the neutral mutation rate. Given this assumption, the following equality holds:

   $$\bar{\omega} \cdot r_{s1} \cdot (T - T_N) + r_{s1} \cdot T_N = Ka_1\,,$$

   in which
   T is the time since the last common ancestor of human/mouse/rat (node A),
   T_N is the time since *ψFXRβ* inactivation (to be estimated),
   $r_{s1} = Ka_1/T$ is the synonymous substitution rate in the lineage 1,
   $\bar{\omega} = \sum_{i=2}^{5} \omega_i / 4$ is the average *Ka*/*Ks* ratio (averaged from all lineages except lineage 1),
   Ka_1 is the nonsynonymous substitutions per nonsynonymous site in the lineage 1.
2. Rearrange the equation above, we have

   $$T_N = T \cdot \frac{\omega_1 - \bar{\omega}}{(1 - \bar{\omega})}\,.$$

   Given $T = 92$ Mya (Hedges 2002), $\omega_1 = 0.6034$, and $\bar{\omega} = 0.2715$, $T_N = 42$ Mya.
3. Due to the small number of species used to estimate T_N, its estimated value should be viewed with caution.



Zebrafish FXR in chr18:17,988,914-18,018,034

Zebrafish FXRβ in chr8:8,073,617-8,086,283

Pufferfish FXR in chrUn:178,182,424-178,188,879

Pufferfish FXRβ in chrUn:214,701,598-214,705,655

Supplementary figure 1. *FXR* and *FXRβ* assignment in zebrafish and pufferfish genomes. Zebrafish mRNA BC092785 is believed to be a *FXR* transcript given the superposition of the human FXR alignment. Zebrafish mRNA DQ017614 is annotated as *FXRβ* mRNA (partial CDS) in GenBank. Despite the strong evidence, the assignment of *FXR* and *FXRβ* in pufferfish is tentative, given its small mRNA set and the early stage of its genome assembly. The genome assemblies used are danRer4 (March 2006) and fr2 (October 2004) for zebrafish (*Danio rerio*) and pufferfish (*Takifugu rubripes*), respectively.

```
Human_FXR      MGSKM-NLIE HSHLPTTDEF SFS------- --ENLFGVLT EQVAGPLGQ- NLEVEPYSQY SNVQFP-QVQ PQI---SSSS YYSNLGFYPQ Q-PEEWYSP-
Mouse_FXR      MVMQFQGLEN PIQISLHHSH RLSGFVPDGM SVKPAKGMLT EHAAGPLGQ- NLDLESYSPY NNVPFP-QVQ PQI---SSSS YYSNLGFYPQ Q-PEDWYSP-
Rat_FXR        M-----NLIG PSHLQATDEF ALS------- --ENLFGVLT EHAAGPLGQ- NLDLESYSPY NNVQFP-QVQ PQI---SSSS YYSNLGFYPQ Q-PEDWYSP-
Dog_FXR        MGSKM-NLIE HSHLPVTEEF SLS------- --DNLFGVLT EQAAGPRGQ- NLDVEPYSQY NNVQFP-QVQ PQI---SSSS YYSNLGFYPQ H-PEEWYSS-
Chicken_FXR    MGSEM-NLIG HPQLATADGF SLA------- EGPHLFGILS EPMSSPVQEA D--VSPYTQY NSVPFP-QVQ PQI---SSPP YYSNLGFYPP Q-HEEWYSP-
Frog_FXR       ---------- ---------- ---------- ---------- ---------- -------SPY NHVQYP-SVH QSMTSSSSSP YHLNSNYYSQ H-AEEWCAN-
Zebrafish_FXR  VGHDV-NVVG PLQIPPNDAF PLS------- ESSHFFDILA EQ-NSPLLQ- DQEVMPFTSY PSMQYT-SVE PSM---SSPS YYSSQHCYSQ YGAEEWYSPS
Human_pFXRb    ---------- ---------- ---------- ------DILP EQISYQLHDT HFKKSPYCQY SIAQFP-PAL QSE---SLXN HFNTYRLDPQ DSDGGQCGF-
Mouse_FXRb     --------MA NTYVATSDGY YLA------- EPTQYYDILP EQFHYQLCDT DFQEPPYCQY STAQFP-PAL QSP---SLQS HFNTHGLDPQ YSGGSWCGL-
Rat_FXRb       --------MA NTYVTTSDGY YLA------- EPTQYYDILP EQLHYQLCDT DFQEPPYSQY STAQFP-PAL QSP---SLQS HFSTYGLEPQ YSGGSWCGL-
Dog_FXRb       --------MA NTYVTTSDGY CLA------- EPVQYYDILP EQINYQLHDT DFQESPYCQY STVQFP-SAL QTQ---SLQS HFSSYSLDPQ F-SGGECGF-
Chicken_FXRb   --------MA NTFVTVPDGY CLA------- EPIQYYDVLP EHINYQLQDT DFQTAPYYQY SSAQIPSPVL QSQ---PSQS HYSAYSLDSQ YTDGQYI-I-
Frog_FXRb      --------MA NSYVTVSDAY CLA------- EPLSYYDVLP DHINYQLPDS EFQTASCCQY TNMAYS-PGL QSP---SSQC HYTSYGLEAA YGDGQYL-L-
Human_LXRa     ---------- ---------- ---------- ---------- ---------- ---------- ----MP---- ---------- ---------- HSAGGTAGV-

Human_FXR      GIYELRRMPA ETLYQGETE- -VAEMP-VTK KPRMGA-SAG RIKGDE--LC VVCGDRA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Mouse_FXR      GIYELRRMPA ETGYQGETE- -VSEMP-VTK KPRMAAASAG RIKGDE--LC VVCGDRA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Rat_FXR        GLYELRRMPT ESVYQGETE- -VSEMP-VTK KPRMAASSAG RIKGDE--LC VVCGDRA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Dog_FXR        GIYELRRMPA ETVYQGEIE- -VAEIP-VTK KARMGA-SAG RIKGDE--LC VVCGDRA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Chicken_FXR    GMYELRRIPS ETFFTRETE- -IMDIP-AAK KPRLGH-STG RMKGEE--LC VVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Frog_FXR       GIYDLKRIPS ENLYSIDTD- -IISLP-ATK KHRVSP-RVG RVKGDE--LC VVCGDNA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Zebrafish_FXR  AMFEMRKGPL DGGFDNELDE SCPVIPTVCK RSRHAG-HSG KSKGEE--LC VVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYK
Human_pFXRb    SSRELNKPTY VVAHDAEDG- ----YP-GIK RSRPTY-SSS RNKGQEE-FC VVCGDKASPS PYHYNALTCE GCKEIPMVKN FKTFLLGFFQ CSIXQNAVYS
Mouse_FXRb     DARESGQSTY VVVHDDEDE- ----FP-GAQ RCRAT--CSL RWKGQDDMLC MVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYS
Rat_FXRb       DTRESSQSTY VVVHDDEDE- ----FP-GTQ RCRPT--CSL RWKGQDE-LC MVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYS
Dog_FXRb       GSYELSKPTF VVDHDAEDG- ----YS-GIK RSSLTH-SSI RLKRQEE-LC VVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYH
Chicken_FXRb   SNCELSKPPF TASHLDDSG- ----FQ-ALK RPRLNH-SSL RLKGQEE-LC VVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYR
Frog_FXRb      STCELSKQTT LMTHGVDEV- ----YP-TMK RPRVSH-ASI RMKGHEE-LC VVCGDKA--S GYHYNALTCE GCK------- ------GFFR RSITKNAVYR
Human_LXRa     GL-EAAEPTA LLTRAEPPS- ----EPTEIR PQKRKKGPAP KMLGNE--LC SVCGDKA--S GFHYNVLSCE GCK------- ------GFFR RSVIKGAHYI

Human_FXR      CKNGGNCVMD MYMRRKCQEC RLRKCKEMGM LAEC----LL TEIQCKSKRL RKNVKQHAD- -QTVNE-DSE GRDLRQVTST TKSCR----- ----------
Mouse_FXR      CKNGGNCVMD MYMRRKCQEC RLRKCREMGM LAEC----LL TEIQCKSKRL RKNVKQHAD- -QTVNEDDSE GRDLRQVTST TKFCR----- ----------
Rat_FXR        CKNGGNCVMD MYMRRKCQDC RLRKCREMGM LAEC----LL TEIQCKSKRL RKNVKQHAD- -QTVNE-DSE GRDLRQVTST TKLCR----- ----------
Dog_FXR        CKNGGNCVMD MYMRRKCQEC RLRKCKEMGM LAECMYTGLL TEIQCKSKRL RKNVKQHAD- -QTINE-DSE GRDLRQVTST TKSCR----- ----------
Chicken_FXR    CKNGGNCEMD MYMRRKCQEC RLRKCKQMGM LAEC----LL TEIQCKSKRL RKNVKQLPD- -QTVNE-DNE GHDMKQVTST TKMYR----- ----------
Frog_FXR       CKNGGNCEMD MYMRRKCQEC RLRKCKQMGM LAEC----LL TEIQCKSKRL RKHAKPQSE- -KSFQE-DID GHETKQVTST TKTNQ----- ----------
Zebrafish_FXR  CKSGGNCEMD MYMRRKCQEC RLRKCKEMGM LAEC----LL TEIQCKSKRL RKNTKASSD- -ESIGDDVVD SRDPKQVVST TKPSK----- ----------
Human_pFXRb    CRNGSHCEMD MYMRRKCQEC RLKKYKAVGM LAEC----LL TEIQCKLKRL QKNFKEKNHF YSNIKV-EEE GVDHSFLSST TRPGK----- ----------
Mouse_FXRb     CKNGGHCEMD MYMRRKCQEC RLKKCKAVGM LAEC----LL TEIQCKSKRL RKNFKHGPAL YPAIQV-EDE GADTKHVSSS TRSGKG---- ----------
Rat_FXRb       CKNGGHCEMD MYMRRKCPEC RLKKCKAVGM LAEC----LL TEIQCKSKRL RKSFKHRPTL SSAIQV-EDE GTDTKHVSST SRSGKGARLF FHTVCPSVSL
Dog_FXRb       CKNGGHCEMD MYMRRKCQEC RLKKCKAVGM LAEC----LL TEIQCKSKRL RKNFKQKNSF YSSIKV-EEE GVD-KLVSST TRSGK----- ----------
Chicken_FXRb   CKNGGHCEMD MYMRRKCQEC RLKKCRAVGM LAEC----LL TEVQCKSKRL RKNFKQKSSF LCNIKL-EDE GVNSKHVSST TRSGK----- ----------
Frog_FXRb      CKNGGHCEMD MYMRRKCQEC RLKKCKAVGM LAEC----LL TEVQCKSKRL RKNCKQNNSM LSNVKV-EDE GSDSRHVSST TKPTK----- ----------
Human_LXRa     CHSGGHCPMD TYMRRKCQEC RLRKCRQAGM REEC----VL SEEQIRLKKL KRQE------ -------EEQ AHATSLPPRA SSPPQ----- ----------

Human_FXR      ---EKTELTP DQQTLLHFIM DSYNKQR--- -------MPQ EITNKI-LKE EFSAEENFLI LTEMATNHVQ VLVEFTKKLP GFQTLDHEDQ IALLKGSAVE
Mouse_FXR      ---EKTELTA DQQTLLDYIM DSYNKQR--- -------MPQ EITNKI-LKE EFSAEENFLI LTEMATSHVQ ILVEFTKKLP GFQTLDHEDQ IALLKGSAVE
Rat_FXR        ---EKTELTV DQQTLLDYIM DSYSKQR--- -------MPQ EITNKI-LKE EFSAEENFLI LTEMATSHVQ ILVEFTKRLP GFQTLDHEDQ IALLKGSAVE
Dog_FXR        ---EKTELTP DQQNLLHYIM DSYSKQR--- -------MPQ EIANKI-LKE EFSAEENFLI LTEMATSHVQ ILVEFTKTLP GFQTLDHEDQ IALLKGSAVE
Chicken_FXR    ---EKVEFTP EQQNLLDYIM DSYSKQQ--- -------IPQ EVSKKL-LHE EFSAEGNFLI LTEMATSHVQ VLVEFTKKLP GFQTLDHEDQ IALLKGSAVE
Frog_FXR       ---ENTELTQ EQMNLLQYVM DSHVKNR--- -------LPQ SLATRLILQE DMGSDDNFVF LTEMATRHVQ ILVEFTKKLP GFQTLDHEDQ IALLKGSAVE
Zebrafish_FXR  ---ENIELSQ DQQALINYIV DAHNKHR--- -------IPQ DMAKKL-LQE QFNAEENFLL LTEMATSHVQ VLVEFTKNIP GFQSLDHEDQ IALLKGSAVE
Human_pFXRb    -IQESMELTE EEHQLINNIV AAHQKYT--- -------IPL EETNLY-LQE HTNPELSFLQ LSETAVLHIR GLMNFTKGLP GFENLANEDQ TALQKGSKTE
Mouse_FXRb     -VQDNMTLTQ EEHRLLNTIV TAHQKSM--- -------IPL GETSKL-LQE GSNPELSFLR LSEVSVLHIQ GLMKFTKGLP GFENLTTEDQ AALQKASKTE
Rat_FXRb       QAQDDMTLTA EERRLLNTIV TAHRKSM--- -------VPV GEISAL-LQE YSNPELSFLR LSEASILHAN WLMKFTKGLP GFENLTAEDQ TALQKESKTE
Dog_FXRb       -IKESVELTQ EEHQLINNIV AAHQKYT--- -------IPL EETKKF-LQK YANPELSFLR LSETVVLHLQ GLIDFTKELP GFENLTIEDQ TALRKGSKTE
Chicken_FXRb   -TVEKVELTP GEHQLLDHIV AAHQKYT--- -------IPL EEARKF-LQE TTSPEESFLH LSETAVVHVQ VLVDFTKRLP GFESLASEDQ IALLKGSTVE
Frog_FXRb      -LSSQPELTA EECKLIDHIV TAHQKCG--- -------IPL DDLKIF-VKE SADPEEIFYH FSEAAVLHVQ AFVEFTKRLP GFEMLDHEDQ IALLKGSTVE
Human_LXRa     ---ILPQLSP EQLGMIEKLV AAQQQCNRRS FSDRLRVTPW PMAPD--PHS REARQQRFAH FTELAIVSVQ EIVDFAKQLP GFLQLSREDQ IALLKTSAIE

Human_FXR      AMFLRSAEIF NKKL------ ---------- ---PSGHSDL LEERIRNS-- ---------- ------GISD EYITPMFSFY KSIGELKMTQ EEYALLTAIV
Mouse_FXR      AMFLRSAEIF NKKL------ ---------- ---PAGHADL LEERIRKS-- ---------- ------GISD EYITPMFSFY KSVGELKMTQ EEYALLTAIV
Rat_FXR        AMFLRSAEIF NKKL------ ---------- ---PAGHADL LEERIRKS-- ---------- ------GISD EYITPMFSFY KSVGELKMTQ EEYALLTAIV
Dog_FXR        AMFLRSAEIF NKKL------ ---------- ---PAGHADL LEERIRKS-- ---------- ------GISD EYITPMFSFY KSVAELKMTQ EEYALLTAIV
Chicken_FXR    AMFLRSAEIF SRKL------ ---------- ---PTGHTVL LEERIRNS-- ---------- ------GISD EFITPMFNFY KSIGELKMTQ EEYALLTAIV
Frog_FXR       AMFLRSAELF NRKL------ ---------- ---LERHTEV LEERIRKS-- ---------- ------GISH DYINPMFHFY KSVGELKMVE EEYALLTAVV
Zebrafish_FXR  AMFLRSAQVF SKKL------ ---------- ---PNGHTEV LEDRIRRS-- ---------- ------GISE EFITPMFNFY KSIGELQMMQ EEHALLTAIT
Human_pFXRb    VIFLHGAQLY SQKQ---SAS ESSVRILNHS DYTPNCHNRS GDRSLICSME KFYNEECPST TLIVFWVLLK NLLXTLFYFY KRMSKLDVTN TEYALLAA-T
Mouse_FXRb     VMFLHVAQLY GGKD---STS GSTMRPAKPS AGTLEVHNPS ADESV-HSPE NFLKEGYPSA PLT---DITK EFIASLSYFY RRMSELHVSD TEYALLTATT
Rat_FXRb       VMFLHVAQLY GGRD---STS GSTVRPAKPS AGTLEVHNHR GDECV-YSSE NFFKEGYPSA TLT---GITR EFIASLSYFY RRMRELNITD TEYALLTATT
Dog_FXRb       VMFLHGAQLY SQKQCLSSAS ESTMRIADHS DHSLNFHNQS DNRNVIYSVE TFHNEDCLPT TLT---GIAE EFITTLFYFY RRMSELNITN IEYALLAATT
Chicken_FXRb   AMLLCSAQIY NQRI--SECQ SSSESHIRRS DHTTCCHVPN LDKN-MYSIQ MSHSEESPTS TTTT--GITE EFITALFYFY RSMGELKVTE TEYALLVATT
Frog_FXRb      AMLLRSAQIY NLPV------ ---------- ---MGCSLQT TEVYFYYVFK IFIKHFSFFS FVI---DLTE EFITPLFKFF RSMGSLNVTE AEYALLSAVT
Human_LXRa     VMLLETSRRY N--------- ---------- ---PGSESIT FLKDFSYNRE DFAKA----- ------GLQV EFINPIFEFS RAMNELQLND AEFALLIAIS

Human_FXR      ILSPDRQYIK DREAVEKLQE PLLDVLQKLC KIHQPENPQH FACLLGRLTE LRTFNHHHAE MLMSWRVNDH KFTPLLCEIW DVQ---
Mouse_FXR      ILSPDRQYIK DREAVEKLQE PLLDVLQKLC KMYQPENPQH FACLLGRLTE LRTFNHHHAE MLMSWRVNDH KFTPLLCEIW DVQ---
Rat_FXR        ILSPDRQYIK DREAVEKLQE PLLDVLQKLC KIYQPENPQH FACLLGRLTE LRTFNHHHAE MLMSWRVNDH KFTPLLCEIW DVQ---
Dog_FXR        ILSPDRQYIK DREAVEKLQE PLLDVLQKLC KIYQPENPQH FACLLGRLTE LRTFNHHHAE MLMSWRVNDH KFTPLLCEIW DVQ---
Chicken_FXR    ILSPDRQYIK DRESVERLQE PLLDILQKFC KLHHPDNPQH FACLLGRLTE LRTFNHHHAE MLMSWRVNDH KFTPLLCEIW DVQ---
Frog_FXR       ILTPDRQYLK DKESVEKLQE TFLHILEKIC KRCHPDNPQH FARLLGRLTE LRTFSHHHAD MLMSWRVNDH KFTPLLCEIW DVQ---
Zebrafish_FXR  ILSPDRPYVK DQQAVERLQE PMLEVLRKIC KLQHPQEPQH FARLLGRLTE LRTLNHHHAE MLESWRMSDH KFNPLLCEIW DVQ---
Human_pFXRb    IVFSDRPCLK NKQYMENLXE PVLQILYKYS KMYHPEDPXH FAHLIWKHTE LRTLNYNHSE ILSTWKTKDP KLATLLSEK- ------
Mouse_FXRb     VLFSDRPCLK NKQHIENLQE PVLQLLFKFS KMYHPEDPQH FAHLIGRLTE LRTLSHSHSE ILRMWKTKDP RLVMLFSEKW DLHSFS
Rat_FXRb       VLFSDRPYLK NKQHVENLQE PVLQLLFKYS KMYHPEDPQH FAHLIGRLTE LRTLSHSHSE ILSTWKTKDP RLVMLFSEKW DLHSL-
Dog_FXRb       VFFSDRPHLK NKRHVENLQE PILHILYKYS KIYHPEDLQH FAHLIGRLTE LRTLNHNYSE ILSTWKAKDP ---------- ------
Chicken_FXRb   VLFSDRPLLR NKRHVEELQE PFLGILYKYS KIHHPEDPQH FARLIGRLTQ LRTLNHTHAE VLVTWRTKDP RLTALLCEVW ELH---
Frog_FXRb      VFFSDRPLLQ NKPHVEKLQE PLLGILHKYS KLYHPEDPQH FARLIGRLTE LRTLNHNHSE VLISWKARDT KLTPLLYGFW NL----
Human_LXRa     IFSADRPNVQ DQLQVERLQH TYVEALHAYV SIHHPHDRLM FPRMLMKLVS LRTLSSVHSE QVFALRLQDK KLPPLLSEIW DVHE--
```

Supplementary figure 2. The FXR and FXRβ multiple sequence alignment. This alignment of FXR and FXRβ peptide sequences was used to produce the FXR-FXRβ phylogeny in Figure 4B. The alignment of sequences in the FASTA format is available at http://pseudogene.org/nr.