Comparative genomics and phylogenetics of the vertebrate *CYP3* family

Dissertation
Zur Erlangung des Grades
“Doktor der Naturwissenschaften”

Am Fachbereich Biologie
der Johannes Gutenberg-Univerisität
in Mainz

Huan Qiu
aus
Jiangxi, China

Mainz
2008

Table of contents

1. Introduction

1. 1 The Cytochrome P450 superfamily

P450 is a superfamily of hemoproteins which have been found in all five kingdoms of life, i.e. in Animalia, Plantae, Fungi, Protista, Archaea, and Eubacteria (Dr. Nelson's Cytochrome P450 Homepage, http://drnelson.utmem.edu/CytochromeP450.html). These proteins were originally identified in 1958 as reduced pigments and named due to their maximum absorption band at 450 nm when binding carbon monoxide, rather than at 420 nm as other hemoproteins (Klingenberg 1958). P450 enzymes participate in the metabolism of both endogenous substances (such as steroids, fatty acids, prostaglandin, and cholesterol) and foreign compounds (including drugs, carcinogens, pollutants, pesticides, etc). These enzymes catalyze monooxygenase reactions, in which the oxygen atoms are inserted into the substrates by activating molecular oxygen and thus introduce hydroxyl groups. These so-called Phase I reactions facilitate the subsequent modifications by phase II enzymes, leading to increased polarity of conjugated substrates and elimination from the body.

While prokaryotic P450 are soluble proteins, eukaryotic P450 are associated with either endoplasmatic reticulum or mitochondrial membrane (Werck-Reichhart and Feyereisen 2000). Although P450 are highly diversified, with sequence identity even below 20% in some cases, their topologies and structures are well conserved. All P450 exhibit a general tertiary structure which consists of several beta-sheet elements at the N-terminus and many alpha-helices in the C-terminal domain designated as helix A–L. The most conserved part is the core of the protein, which is composed of two motifs on the proximal side of heme (heme-binding loop-containing motif: Phe-X-X-Gly-X-Arg-X-Cys-X-Gly and helix K-containing motif: Glu-X-X-Arg) and helix I on the distal side of heme that contains the Ala/Gly-Gly-X-Asp/Glu-Thr-Thr/Ser motif (Werck-Reichhart and Feyereisen 2000).

CYP families are defined by at least 40% amino acid sequence identity, whereas the corresponding number for subfamilies is 55% (Nelson et al. 1996). In the human genome there are 57 P450 encoding genes and 58 related pseudogenes which belong to 18 different families (Nelson et al. 2004). P450 involved in xenobiotics metabolism include members of CYP1, CYP2 and CYP3 families. These proteins are mainly expressed in the liver and in the small intestine and they are believed to be responsible for approximately 80% of oxidative drug metabolism and about 50% of the overall elimination of commonly used drugs (Wilkinson 2005). Among them, CYP3A family is the most important one due to its abundant expression and unusually wide substrate spectrum. CYP3A4 alone accounts for approximately 30% of the total P450 expression in liver and up to 80% in the small intestine (Paine et al. 2006; Shimada et al. 1994), and metabolizes more than half of the clinically used drugs (Evans and Relling 1999). Thus, the investigation of CYP3A is of clinical relevance.

1.2 Human *CYP3A* family

1.2.1 Structure of human *CYP3A* genes and locus

Human *CYP3A4*, *CYP3A5*, *CYP3A7*, and *CYP3A43* form a gene cluster of ~250 kb on chromosome 7q21-22.1 (Finta and Zaphiropoulos 2000; Gellner et al. 2001). The locus is enriched in repeat elements which account for approximately half of its size (Fig. 1). All *CYP3A* are 13-exon genes comprising 26~38 kb with the intergenic distances between each other ranging from 23 to 44 kb (Table 1). *CYP3A5*, *CYP3A7* and *CYP3A4* are arranged in a head-to-tail manner on one strand and in a head-to-head orientation to *CYP3A43,* which is located on the other strand. These four genes are supposed to be derived from successive local gene duplication events (Finta and Zaphiropoulos 2000) and share 82~94% and 71~88% identity at coding sequence and protein levels, respectively. While *CYP3A5* and *CYP3A43* are located in the 13-exon-containing duplicative fragments, *CYP3A4* and *CYP3A7* result from duplication of fragments consisting of at least 15 exons (the 13 canonical exons and two detritus exons downstream of each gene) (Finta and Zaphiropoulos 2000). *CYP3A7* is most closely related to *CYP3A4,* as judged from their sequence similarity and relative physical positions in the locus. *CYP3A5* is the second closest relative of *CYP3A4* and *CYP3A43* the least similar one.

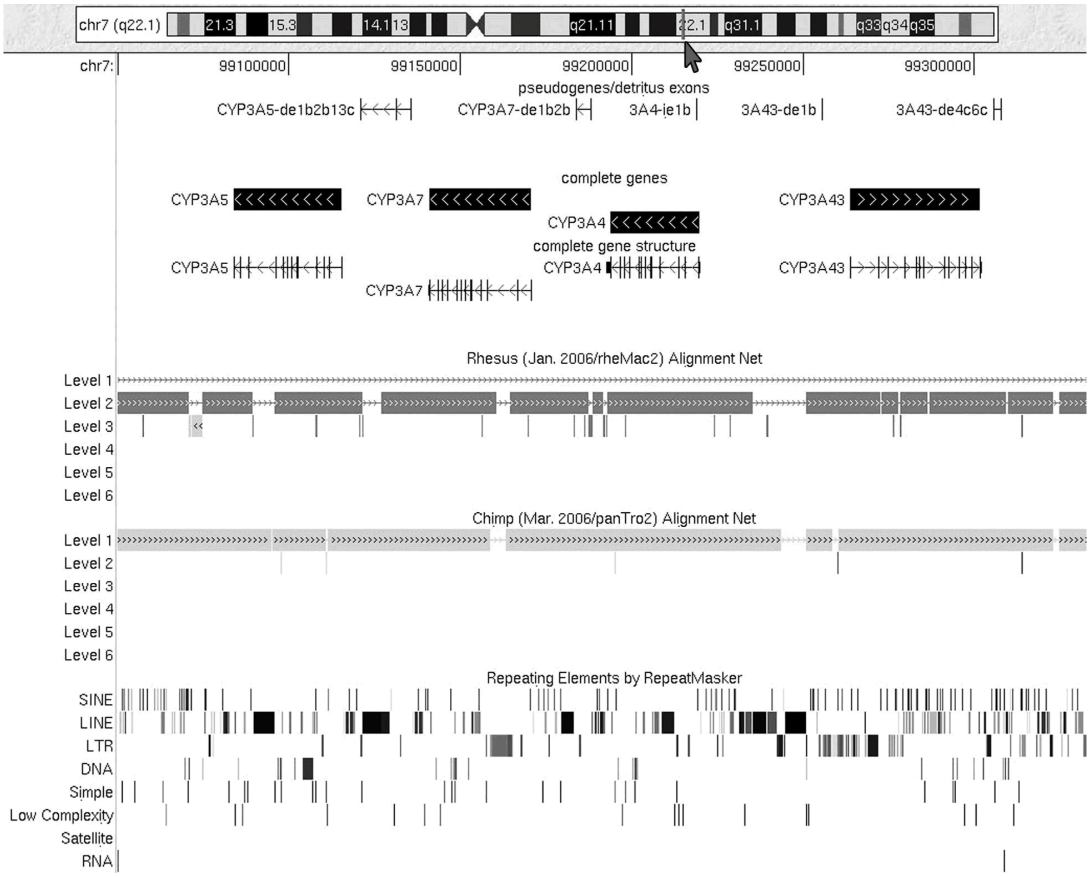


Fig. 1. Structure of the human *CYP3A* locus. From top to bottom: The position of the *CYP3A* locus on human chromosome 7 (genome assembly: hg18) is indicated by the arrow below the chromosome ideogram. Detailed genomic locations are indicated by numbers along the line beneath the chromosome ideogram. Genes, pseudogenes and detailed complete gene structure annotation is displayed in their corresponding tracks. Vertical and horizontal lines represent exons and introns, respectively. Arrows indicate orientation of genes. The Chimp and Rhesus Alignment Net tracks show ortholog regions in chimp and rhesus genome, respectively. Repetitive elements are indicated by vertical lines and black boxes in Repeating Elements by RepeatMasker track. The annotation of human *CYP3A* locus was uploaded to and displayed by UCSU genome browser (Hinrichs et al. 2006).

A number of pseudogenes (*CYP3A5-de1b2b*, *CYP3A5-de13c*, *CYP3A7-de1b2b*, *CYP3A4-ie1b*, *CYP3A43-de1b* and *CYP3A43-de4c6c*) consisting of detritus exons are also found in the locus (Dr. Nelson's Cytochrome P450 Homepage, Fig. 1). Both *CYP3A5P1* (*CYP3A5-de1b2b* and *CYP3A5-de13c*) and *CYP3A5P2* (*CYP3A7-de1b2b*) are assumed to have arisen from a disrupted ancient *CYP3A* gene (Finta and Zaphiropoulos 2000). The lost ancient *CYP3A* genes was a chimerical gene with the first exon being identical to extant *CYP3A5* coding exon 1 and the other two exons displaying strong sequence similarity to *CYP3A7* exon 2 and exon 13. The creation of this chimerical *CYP3A* gene might be due to a recombination event between ancient *CYP3A5* and *CYP3A4*/*CYP3A7*-like gene with the breakpoint located in their first introns. *CYP3A5P2*, located downstream of *CYP3A4*, is the duplicative product of the first two exons of *CYP3A5P1* or vice versa (Finta and Zaphiropoulos 2000).

Human *CYP3A* encode proteins consisting of 502-504 amino acids (Table 1). All human *CYP3A* transcripts contain short 5'UTR of approximately 100 bp and 3'UTR of 111 to 1152 bp. The known longest 3'UTR, 1152 bp, has been identified in *CYP3A4*, and it is due to the alternative use of a second polyadenylation signal downstream of this gene. However, this long transcript accounts for only one tenth of the expression level of *CYP3A4* transcripts containing the 457-bp 3'UTR (Bork et al. 1989). The significance of the existence of two 3'-UTR in 3A4 is unknown. *CYP3A43* is considered a pseudogene, based on a low level of mostly aberrant transcripts, although bacteria-expressed protein exhibits some activity (Daly 2006). *CYP3A5* is also aberrantly spliced in some individuals, the percentage of whom is population-specific (see section 1.2.3). Interestingly, the last two exons of the pseudogene *CYP3A5P1* can be transcribed and spliced into wildtype *CYP3A7* transcripts. The resulting transcript is expressed in multiple tissues and encodes an enzyme which differs functionally from the wildtype *CYP3A7* (Finta and Zaphiropoulos 2000; Rodriguez-Antona et al. 2005). Even more strikingly, trans-splicing events were detected among members *CYP3A* family with the first exon of *CYP3A43* spliced to either *CYP3A4* or *CYP3A7* downstream exons. Due to their extreme low expression level, the functional consequences of these chimerical *CYP3A* transcripts are unlikely to be significant (Finta and Zaphiropoulos 2002).

Table 1. Statistics of human *CYP3A* genes and locus

Genes	*CYP3A5*	*CYP3A5P1**	*CYP3A7*	*CYP3A5P2**	*CYP3A4*	*CYP3A5P3**	*CYP3A43*
Protein length (aa)	502		503		503		504
Gene length (bp)	31592	25603$^{\&}$	29594	23039$^{\&}$	25949	44034$^{\&}$	37886
5'UTR (bp)	87		105		104		103
3'UTR (bp)	111		463		457/1152		549
Number of exons	13	3	13	2	13	1	13
Genomic location (chr7)	99083864-99115455		99141059-99170652		99193692-99219640		99263675-99301560
Repeats content	0.452266	0.597703	0.410759	0.425322	0.35038	0.694759	0.055271
Strand	-		-		-		+

All data are based on human genome assembly (hg18). $^{\&}$ length of intergenic regions; * pseudogenes within the intergenic regions.

1.2.2 Substrates of human CYP3A

Besides steroid hormones, cholesterol and other endogenous substrates, CYP3A4 metabolizes at least every-second drug currently in use (Wilkinson 2005). In particular, CYP3A4 is capable of accommodating large molecules such as cyclosporine and bromocriptine and exhibits non-Michaelis-Menten kinetics toward some substrates (Atkins 2005). These characteristics are speculated to be due to either multiple ligand-binding sites within the protein tertiary structure (He et al. 2003; Kenworthy et al. 2001) or to kinetic changes of CYP3A4 conformation when bound to different ligands (Davydov et al. 2003; Johnson and Stout 2005; Koley et al. 1997). One of the consequences of ligand promiscuity of CYP3A4 is its frequent involvement in clinically relevant drug-drug interactions. Another factor which complicates therapies with CYP3A4 drug substrates is the unpredictable individual CYP3A4 expression level in the liver and in the small intestine, which is assumed to be inherited (Ozdemir et al. 2000). This variability may be further enhanced by CYP3A4 induction or inhibition by certain drugs and dietary constituents and contribute to drug interactions involving this isozyme.

The known substrate spectra for CYP3A5 and CYP3A7 are generally smaller in comparison to CYP3A4 (Daly 2006). For most substrates, CYP3A5 and CYP3A7 generally display lower metabolic capability compared to CYP3A4 (Williams et al. 2002), but there are exceptions. For example, the intrinsic clearance for vincristine is 9- to 14-fold higher for CYP3A5 than for CYP3A4 (Dennison et al. 2006). 1’-hydroxylation of alprozolam is preferentially catalyzed by CYP3A5, with V_{max} of CYP3A5 being at least two fold higher than that of CYP3A4 (Galetin et al. 2004; Williams et al. 2002). Although it is still debated, most studies have shown higher rates of formation of 1’-hydroxymidazolam from midazolam by CYP3A5 than by CYP3A4 (Galetin et al. 2004; Huang et al. 2004). Tacrolimus has been reported to be metabolized by CYP3A5 with a catalytic efficiency 64% higher than that of CYP3A4 (Kamdem et al. 2005). CYP3A7 displays higher catalytic activity towards retinoic acid isomers in comparison to CYP3A4 and CYP3A5 (Chen et al. 2000; Marill et al. 2000). It is supposed to account for up to 80% of the retinoic acid metabolism in individuals expressing CYP3A7 post-natally (Burk et al. 2002). Compared to CYP3A4 and CYP3A5, CYP3A7 also shows higher 16-hydroxylation activity towards estrogen (Lee et al. 2003) and dehydroepiandrosterone (Kitada et al. 1987).

1.2.3 Variability in the expression of human *CYP3A*

In agreement with their well-known role in the detoxification of exogenous compounds, CYP3A isozymes are most abundantly expressed in the human liver and small intestine, with the expression level in the former organ accounting for 30~60% of the total CYP protein (Shimada et al. 1994). Expression at protein level *in vivo* has been demonstrated for CYP3A4, CYP3A5, and CYP3A7, but not CYP3A43. The level of CYP3A proteins correlate with the corresponding mRNA expression level whereas the effects of posttranscriptional regulation are assumed to be negligible. Among the CYP3A family, CYP3A4 is the most abundant hepatic CYP3A isoform in adults. The individual expression of CYP3A4 varies up to 90 fold (Lamba et al. 2002). The reasons for large inter-individual variation in CYP3A4 expression are still incompletely understood. Since no allelic variants with major effects on CYP3A4 expression have been identified, it could be due to the combined effects of large number of minor variants which have effects on CYP3A4 expression. However, the frequency of most allelic variants is too low to explain the variability in CYP3A4 expression. Therefore alternative factors such as the individual exposure to CYP3A4 inducers and inhibitors have been proposed to be responsible for the large part of the variability. In addition, influence on the CYP3A expression by genetic variants beyond *CYP3A* locus is also possible (Plant 2007; Wojnowski 2004).

In contrast to the unimodal expression of CYP3A4, the distribution of the intestinal and hepatic CYP3A5 expression is bimodal. High expression of CYP3A5 is limited to a part of a given population (~70% Africans, ~30% Asians, and ~10% Central Europeans) (Burk and Wojnowski 2004). While CYP3A5 generally contributes 10~20% of total hepatic CYP3A proteins, its expression level in some cases is comparable to, or even exceeds, that of CYP3A4 (Daly 2006). The *CYP3A5*3/*1* gene variant leads to polymorphic CYP3A5 expression and it is common to all world populations investigated thus far (Kuehl et al. 2001). The low expression allele (*CYP3A5*3*) results in an alternative splice acceptor site and the inclusion of an extra mini-exon into the wildtype transcript. These aberrantly spliced transcripts undergo rapid nonsense-mediated degradation leading to low expression of CYP3A5 (Busi and Cresteil 2005). In Africans, the expression of CYP3A5 is also diminished in the carriers of *CYP3A5*6* and *CYP3A5*7* alleles (Hustert et al. 2001; Kuehl et al. 2001). High expression of CYP3A5 has been found only in the carriers of *CYP3A5*1* alleles. *CYP3A5* expression is bimodal also in the kidney (Haehner et al, 1996), where it constitutes the predominant form of *CYP3A* (Koch et al. 2002). Furthermore, polymorphic

CYP3A5 expression in the kidney has been implicated in hypertension (Givens et al. 2003), although the evidence is still inconclusive (Wojnowski and Kamdem 2006).

Although CYP3A4 and CYP3A7 are most closely related of all four CYP3A members, these two CYP3A isoforms' expression is temporally mutually exclusive in most individuals (Lacroix et al. 1997). CYP3A7 is predominantly expressed in the fetal liver (Bieche et al. 2007; Leeder et al. 2005), where it accounts for more than 30~50 % of total CYP (Shimada et al. 1996) and may protect the fetus from the toxicity of accumulated dehydroepiandrosterone 3-sulfate (DHEA-S) (Kitada et al. 1987; Kitada et al. 1985). Although first regarded as fetal liver-specific, CYP3A7 was later found to be expressed in about 20% adult livers in Europeans (Koch et al. 2002), where it accounts on average for 24% of the total CYP3A (Sim et al. 2005). Polymorphical CYP3A7 expression was also found in the small intestine (Burk et al. 2002). Two thirds of the CYP3A7 "high expressers" are accounted for by the alleles *CYP3A7*1C* and *CYP3A71*B*. (Burk et al. 2002). The former is due to a gene conversion between *CYP3A4* and *CYP3A7* which replaced a stretch of *CYP3A7* promoter sequence with the corresponding part of *CYP3A4* which contains a functional ER6 element (Kuehl et al. 2001). Expression of CYP3A7 has also been reported in endometrium and placenta with putative functions in the maintainance of proper progesterone level during pregnancy (Schuetz et al. 1993). The physiological significance of CYP3A7 expression in several other organs, such as adrenal gland, prostate and kidney (Bieche et al. 2007; Koch et al. 2002) is unclear.

1.2.4 Nuclear receptors and *CYP3A* regulation

Nuclear receptors are a family of structurally related transcription factors activated upon binding of ligands, such as steroid hormones, vitamins, fatty acids and xenobiotic compounds. Many activated nuclear receptors form homodimers or heterodimers (in the latter case with the ubiquitous partner, retinoid X receptor, RXR), which bind to specific DNA response elements consisting of two 6-nucleotide half sites in various relative orientations and separated by spacers of variable length (Fig. 2). The binding of a nuclear receptor to a responsive DNA element initiates the transcription of the associated gene. Due to the large number and ligand diversity, nuclear receptors control a variety of development, homeostatic and xenobiotic response processes.

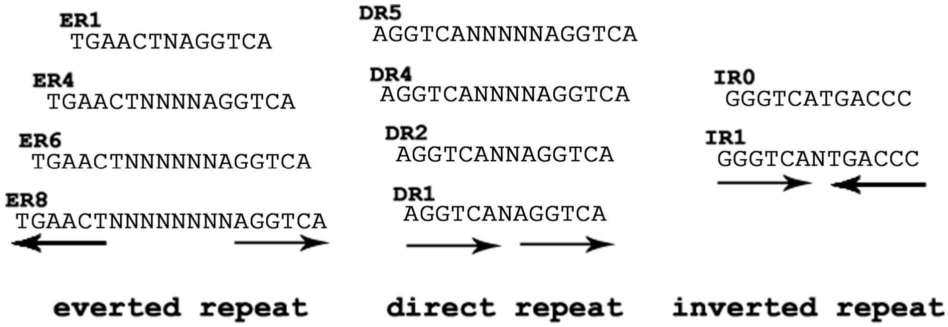


Fig. 2. Examples of nuclear receptor binding sites. Half sites are indicated by arrows. The binding sites are named after the relative orientation of the two half sites and the length of intervening spacer. For example, an everted repeat (ER6) element is composed of two everted AGGTCA hexamers separated by a six-nucleotide spacer.

Most nuclear receptors contain a DNA-binding domain in the N-terminal domain which recognizes the specific DNA target sequence, and a ligand-binding domain in the C-terminus. Four categories of nuclear receptors have been proposed based on the mode of activation: (1) nuclear receptors which form homodimers interacting with inverted repeats, (2) receptors that form heterodimers with RXR and mostly bind to everted repeats (3) orphan nuclear receptors which bind to direct repeats as homodimers and (4) single half-site binding receptors (Mangelsdorf et al. 1995; Olefsky 2001)

There are 48 nuclear receptors-coding genes and 4 related pseudogenes in the human genome (Zhang et al. 2008). Several nuclear receptors have been shown to be involved in the transcriptional regulation of *CYP3A* genes. These include the most investigated xenosensors, i.e. pregnane X receptor (PXR/NR1I2) and constitutive androstane receptor (CAR/NR1I3) (Goodwin et al. 2002). Both PXR and CAR are implicated in the regulation of the constitutive expression of *CYP3A4* and mediate transcriptional activation of this gene (Goodwin et al. 2002). The induction of CYP3A5, both in liver and in the small intestine, is regulated by PXR and CAR as well (Burk et al. 2004). PXR was also shown to mediate the induction of *CYP3A7* in *CYP3A7*1C* allele carriers (Burk et al. 2002). The induction of *CYP3A4* by PXR and CAR is also modulated by a third nuclear receptor, hepatocyte nuclear factor-4-α (HNF4α/NR2A1) (Liu et al. 2008; Tirona et al. 2003), which is crucial in the regulation of *CYP3A4* basal expression in the HepG2 cell line (Matsumura et al. 2004) and in the small intestine (Tegude et al. 2007). Interestingly, HNF4α exerts repressive effects on *CYP3A4* induction by PXR under certain circumstances (Liu et al. 2008). In addition to PXR, farnesoid X receptor (FXR/NR1H4) is also implicated in *CYP3A4* induction by some bile acids (Gnerre et al. 2004; Staudinger et al. 2001). *CYP3A4* is also inducible by vitamin D3 and its derivatives via the vitamin D receptor (VDR/NR1I1) (Thummel et al. 2001). Furthermore, *CYP3A5* has been reported to undergo induction by glucocorticoids, which is mediated through glucocorticoid receptor (GR/NR3C1) (Hukkanen et al. 2003).

1.2.5 Regulatory DNA elements in *CYP3A* promoters

Due to the clinical relevance of *CYP3A4*, its promoter has been intensively investigated. The *CYP3A4* 5'-promoter region contains at least three core modules (Fig. 3) contributing to the induction and constitutive expression of *CYP3A4* (Martinez-Jimenez et al. 2007). The first module is the ER6-containing proximal promoter region located around -160 bp upstream of the *CYP3A4* transcriptional start site (Barwick et al. 1996). The ER6 element has been shown to bind PXR, CAR, as well as VDR, and it is critical for both basal and inducible expression of *CYP3A4* (Goodwin et al. 2002; Thompson et al. 2002; Xie et al. 2000). The second module is called Xenobiotic-Responsive Enhancer Module (XREM), and it is located between -7.8 to -7.2 kb with respect to transcriptional start site (Goodwin et al. 1999). A cluster of elements responsive to PXR, CAR, VDR, FXR, and HNF-4α has been identified within this region. The dNR1 (DR3) and dNR2 (ER6) mediate the transcriptional activation of *CYP3A4* by PXR (Goodwin et al. 1999). The DR3, but not ER6, was later found to be responsive to CAR and VDR (Goodwin et al. 2002; Thompson et al. 2002). Further investigation showed that the XREM-mediated xenobiotic induction in response to PXR and CAR was affected by HNF4α through a nearby binding site (Tirona et al. 2003). Moreover, two more functional elements (an ER8 element 55 bp downstream of dNR2 and a DR1 overlapping with the 5' half site of dNR1) were subsequently identified to be responsible for *CYP3A4* induction by bile acid via FXR (Gnerre et al. 2004). FXR is also able to bind to the PXR- and CAR-responsive dNR1 (Gnerre et al. 2004). The third core module is the Constitutive Liver Enhancer Module (CLEM) in a region form -11.4 kb to -10.5 kb (Matsumura et al. 2004). It contains a number of DNA elements essential for the maximal enhancer activity: two HNF-4alpha binding sites, one HNF-1 binding site, three E-boxes that interact with SUF1, and an AP1 binding site (Matsumura et al. 2004). More recently, a PXR-binding ER6 site within CLEM was found to be required for the maximal induction of *CYP3A4* by rifampin (Liu et al. 2008).

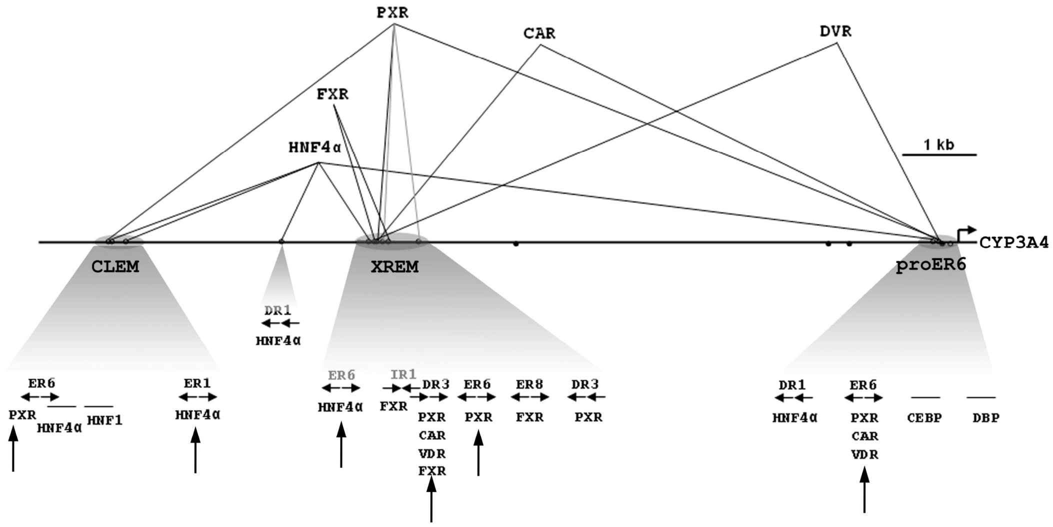


Fig. 3. Known regulatory elements within 13 kb of the *CYP3A4* promoter. The black bar in the middle of figure represents the *CYP3A4* promoter. All DNA elements that interact with nuclear receptors are shown in detail. Black dots beneath the bar indicate C/EBP binding sites. Arrows indicate the nuclear receptor binding sites detectable by NHR-scan (Sandelin and Wasserman 2005) (see section 4.6).

Several additional transcription factor binding sites have been identified outside of these three core modules. Three CCAAT-enhancer-binding proteins (C/EBP) response elements (-121, -1393 and -1659 bp) are involved in the constitutive regulation of *CYP3A4* by C/EBP (Rodriguez-Antona et al. 2003). An additional enhancer site containing 4 C/EBP response elements has been identified at -5.95 kb and may be involved in the *CYP3A4* regulation under pathophysiological conditions (Martinez-Jimenez et al. 2005). Two DR1 elements (-210 and -9040 bp) which bind to HNF4α were found to be crucial for *CYP3A4* basal expression in the small intestine (Tegude et al. 2007). More recently, a binding site for the circadian transcriptional factor, D-site-binding protein (BDP) at -24 bp, was reported to be relevant to the 24-hour rhythmic expression of *CYP3A4* (Takiguchi et al. 2007).

The promoter regions of *CYP3A7* and *CYP3A5* are less well understood. The -8.8 kb 5' upstream region of *CYP3A7* shares ~90% identity with that of *CYP3A4* (Bertilsson et al. 2001). However, the proximal ER6 of *CYP3A7* is mutated, with one nucleotide substitution in its 5' half site and another within the 6 bp spacer in comparison with *CYP3A4*. These two substitutions have been shown to abolish PXR- and CAR-dependent transcriptional activation of *CYP3A7* by reducing the binding of ER6 to PXR and CAR (Burk et al. 2002). Loss of *CYP3A7* induction in response to vitamin D3 has been also attributed to these ER6 mutations (Hara et al. 2004). On the other hand, the distal enhancer XREM is well conserved between *CYP3A4* and *CYP3A7* and was shown to mediate *CYP3A7* induction by PXR and CAR (Bertilsson et al. 2001). A single nucleotide difference from *CYP3A4* in a nuclear factor kappa B-like element at minus ~2.3 kb of *CYP3A7* promoter has been shown to confer the expression of *CYP3A7* in human HepG2 cells, and perhaps in the fetal liver (Saito et al. 2001). The *CYP3A7* distal promoter contains no enhancer corresponding to *CYP3A4* CLEM, since promoter regions beyond -8.8 kb of *CYP3A7* and *CYP3A4* are totally different at sequence level. For *CYP3A5*, the proximal ER6 motif, homologous to that of *CYP3A4* and *CYP3A7*, has been shown to mediate *CYP3A5* induction by PXR and CAR (Burk et al. 2004). A functional binding site for GR was characterized in the 5' region of *CYP3A5P1* (Schuetz et al. 1996), which was mistaken as the authentic *CYP3A5* promoter due to the high sequence similarity. The function of the corresponding putative *CYP3A5* GR response element has not yet been characterized.

1.3 *CYP3* evolution

Vertebrate *CYP3* and *CYP5*, together with insect *CYP6*/*9*/*28* and nematode *CYP13*/*25*, belong to the *CYP3* CLAN which may have derived from a single ancestor existing before protostome-deuterostome divergence 680 million years ago (Nelson 1998). Vertebrate *CYP3* family displays dramatic inter-species variation in gene number which may have been caused by the exposure to diverse xenobiotic substances (McArthur et al. 2003; Nelson et al. 2004; Williams et al. 2004a). Likewise, the *CYP6/9/28* from insect (e.g., *Drosophila melanogaster, Anopheles gambiae* and *Apis mellifera*) (Claudianos et al. 2006; Feyereisen 2006; Tijet et al. 2001) (The insect P450 site: http://p450.sophia.inra.fr) and *CYP13/25* from nematode (Dr. Nelson's Cytochrome P450 Homepage) underwent multiple independent duplication events as well. The proteins encoded by most *CYP3* CLAN genes, both form vertebrates and insects, have been shown to be largely involved in the metabolism of xenobiotics, such as drugs, insecticides (Carino et al. 1994; Korytko and Scott 1998), plants (Danielson et al. 1997; Li et al. 2004) and pollutants (Korytko et al. 2000; Stevens et al. 2000). One known exception is the vertebrate *CYP5A* which is maintained as a one-copy gene in all species studied. Rather than metabolizing xenobiotics, CYP5A1 (thromboxane synthase) processes prostaglandin H2 into thromboxane A2 and has been implicated in human cardiovascular diseases involving platelet aggregation (Shen and Tai 1998). The contrast between *CYP5A* and the other genes of the *CYP3* CLAN is in agreement with the recent classification of vertebrate P450 genes into two categories:

Phylogenetically stable genes whose copy number remain the same across diverse species, and phylogenetically unstable genes, the copy of which varies among species. The former genes are generally found to encode enzymes with functions in core development and biochemical pathways, i.e. those metabolizing endogenous substances (e.g., steroid and retinoid hormones), whereas the later ones respond to changes in xenobiotic exposure and catalyze exogenous substrates (Thomas 2007). Thus, *CYP3/6/9/28/13/25* and *CYP5A1* belong to different categories with regard to their phylogenetic stability and substrates, though all may be derived from one *CYP3* CLAN ancestor.

Vertebrate *CYP3* includes four subfamilies: *CYP3A*, *CYP3B*, *CYP3C*, and *CYP3D* (Corley-Smith et al. 2006; Nelson 2003). While *CYP3A* genes have been found in all vertebrates studied to date, *CYP3B-D* are likely to be Clupeocephala specific, since no members of these subfamilies have been identified in other species. More recently, *CYP3* sequences form tunicate species have been also reported, which display a 13-exon gene structure similar to vertebrate *CYP3* genes (Verslycke et al. 2006). For primates, several *CYP3A* coding sequences from Old World Monkeys (green monkey, cynomolgus monkey, rhesus and Japanese monkey) and New World Monkeys (squirrel monkey) are deposited in Genbank database. A number of *CYP3A* Expressed Sequence Tag (EST) sequences are available for orangutan and rhesus (Magness et al. 2005). Besides human, only *CYP3A* in the chimpanzee have been studied on the genomic scale (Williams et al. 2004a). In comparison to humans, chimpanzees have an additional gene (*CYP3A67*), which is located between *CYP3A5* and *CYP3A7* and shares a sister relationship to both human and chimpanzee *CYP3A7* (Williams et al. 2004a). In humans, strong positive selection signal has been reported in the *CYP3A* locus in non-African human populations (Schirmer et al. 2006; Thompson et al. 2004). *CYP3A* locus in these ethnic groups contains low numbers of long-range haplotypes, consistent with their rapid expansion due to positive selection. Salt-water homeostasis (Thompson et al. 2004) and rickets (Schirmer et al. 2006) have been proposed as the underlying selection factors, consistent with the expression of CYP3A5 in the kidney and with the activity of CYP3A4 towards vitamin D, respectively.

1.4 Objectives of this study

Due to the clinical significance of CYP3A, a better understanding of CYP3A function and regulation is extremely important both for the development of future drugs and for the safe application of the existing CYP3A drug substrates. Most rapid progress has been achieved in the field of CYP3A transcriptional regulation (Burk and Wojnowski 2004). On the other hand, it is still incompletely understood what makes a drug a CYP3A4 substrate and which gene variants determine the individual CYP3A4 expression level. Gene function and regulation is increasingly studied using techniques of population genetics and phylogenomics, enabled by the ever-growing number of genome sequencing and genotyping efforts. Previous phylogenetic analyses of *CYP3* have been mainly performed on cDNA or protein sequences (McArthur et al. 2003; Williams et al. 2004a). This had certain, unavoidable limitations, since, in addition to true relatedness, similarities among cDNA sequences are strongly affected by species-specific selection and by genomic recombination events. Moreover, due to limited sampling, *CYP3* evolution in ancient amniota and fish was largely unknown.

Whole genome shotgun sequencing of a rapidly increasing number of species makes possible addressing the evolution of gene families by comparative genomic approaches. Genomic sequences from 16 vertebrate species were used in the present work to reconstruct the genomic evolution of *CYP3A* and of the related subfamilies *CYP3B-D* during the last ~450 million years. This was complemented by the investigation of gene conversion events and a phylogenetic analysis. Special emphasis was put on the study of ~65 million years of primate *CYP3A* evolution, to facilitate the clarification of the paralogous and orthologous relationships among the *CYP3A* genes in these species. Based upon this work, the evolutionary regime that has shaped the contemporary primate CYP3A protein coding sequence was investigated using the ratio of non-synonymous (amino acid altering) and synonymous (silent) substitution rates (= dN/dS = Ka/Ks = ω). A comprehensive analysis of primate *CYP3A* promoter sequences was also conducted, to establish framework for further functional characterization of *CYP3A* regulatory elements.

.

2. Methods and Materials

2.1 Sequence source

Genomic assemblies of 16 species were accessed through either UCSC (Hinrichs et al. 2006) or ENSEMBL (Hubbard et al. 2007) genome browsers or NCBI Genbank database. The investigated genomes were those of human (*Homo sapiens*, hg18), chimpanzee (*Pan troglodytes*, panTro2), rhesus (*Macaca mulatta*, rheMac2), mouse (*Mus musculus*, mm8), rat (*Rattus norvegicus*, rn4), dog (*Canis familiaris*, canFam2), horse (*Equus caballus*, Equus1.0), opossum (*Monodelphis domestica*, monDom4), platypus (*Ornithorhynchus anatinus*, oaNa5), chicken (*Gallus gallus*, galGal3), frog (*Xenopus tropicalis*, xenTro2), zebrafish (*Danio rerio*, danRer4), fugu (*Takifugu rubripes*, fr1), green spotted pufferfish (*Tetraodon nigroviridis*, tetNeig1), medaka (*Oryzias latipes*, HdrR) and stickleback (*Gasterosteus aculeatus*, gasAcu1). In addition, we analyzed the available partial olive baboon (*Papio anubis*) *CYP3A* locus sequence from High Throughput Genomic (HTG) Sequences (Genbank No. AC141417.16). Novel *CYP3* in the above mentioned genomes were searched by using either Basic Local Alignment and Search Tool (BLAST) from NCBI, BLAST-Like Alignment Tool (BLAT) from UCSC or ENSEMBL genome browser. The acquired *CYP3* sequences together with known *CYP3* from Dr. Nelson's Cytochrome P450 Homepage were mapped to their corresponding whole genome assembly, if available, by BLAT. New *CYP3B-D* partial sequences were identified by searching the NCBI EST database via TBLASTN from NCBI.

Partial sequences of the orangutan (*Pongo pygmaeus*) and green anole (*Anolis carolinensis*) *CYP3A* loci were obtained by assembling Whole Genome Shotgun (WGS) sequences. The sequences were obtained by BLAST of known *CYP3A* mRNA sequences to the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). The traces were downloaded, quality clipped and assembled in GAP4 (Bonfield et al. 1995). All assemblies were manually edited and checked for consistency by distance and orientation of mated read pairs. In regions not covered by WGS sequences (sequence gaps), the order of contigs was confirmed by at least two mated-read pairs. The genomic assemblies of platypus and baboon *CYP3A* loci (see above) were also checked and confirmed by this approach.

Marmoset (*Callithrix jacchus*) and galago (*Otolemur garnettii*) *CYP3A*-containing BAC clones were isolated by screening the CHORI-259 and CHORI-256 Bacterial Artificial Chromosome (BAC) libraries, respectively (http://bacpac.chori.org/). The probe comprised 141 bp *CYP3A* exon 12 sequence amplified from marmoset genomic DNA. Shotgun sequencing of the clones CH259-272B6, CH259-48H24, CH256-186P19 and CH256-241K21 was done using dye terminator chemistry and ABI3730 DNA sequencers (Applied Biosystems). Chimpanzee BAC clones (CH251-35M15, CH251-400N23, CH251-373D9, CH251-171D20, CH251-506N22) were obtained from Chimpanzee BAC libraries (CHORI-251) in Children's Hospital Oakland Research Institute (CHORI).

2.2 Phylogeny construction

107 CYP3 amino acid sequences from 31 species were aligned using ClustalX (Jeanmougin et al. 1998) and manually edited to optimize the alignment. Phylogenetic relationships of these CYP3 sequences were reconstructed using Bayesian and Maximal likelihood methods under JTT+G+I substitution model suggested by Prottest v1.3 (Abascal et al. 2005). Bayesian inference was performed using program MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), which estimates posterior probabilities of clade support using Metropolis-coupled Monte Carlo-Markov Chain method (MC^3). The posterior probabilities were estimated using uninformative prior probabilities with inclusion of unequal amino acid frequencies. Two parallel runs, each with four chains, were run for 1 million generations. For each run, three chains were heated and one was cold with a temperate parameter of 0.20. Rate variation across sites was approximated using a four-category gamma distribution and the proportion of invariable sites estimated from the data assuming a uniform prior distribution. Trees were sampled every 100 generations and following a burn-in of 2500 generations. A 95% majority rule consensus tree was generated to calculate posterior probability values. Maximum likelihood analysis was done using PHYML (Guindon and Gascuel 2003; Guindon et al. 2005). The initial tree was determined by Neighbor-joining (BIONJ). Rate variation across sites was approximated using a four-category gamma distribution. The tree topology, branch lengths, gamma-shape parameter, and proportion of invariable sites were optimized by the software during the run. Branch supports were estimated from 1000 PHYML bootstrap replicates.

A rooted phylogenetic tree of 28 primate genes (Fig. 10) based on the full-length coding cDNA and a phylogeny of 20 primate *CYP3A* promoters sequences (Fig. 18A) were reconstructed by Bayesian inference with HKY+G substitution model following the procedure described above for protein sequences analysis. Comparison of tree topologies was performed using Kishino-Hasegawa test (Kishino and Hasegawa 1989) implemented by TREE-PUZZLE 5.2 (Schmidt et al. 2002). To verify the primate *CYP3A* phylogeny, rare genomic changes (RGCs) were investigated in the 20 *CYP3A* genes (Fig. 11) with complete genomic sequences (Appendix Table 8.1). The genomic sequences were aligned using Multi-LAGAN (Brudno et al. 2003) and shared orthologous intronic retroposed elements and random indels (minimum 2 nucleotides) were screened as described (Kriegs et al. 2006). The presence/absence of these elements was then mapped on the topology of a sequence-based phylogenetic tree (Fig. 11).

2.3 Relative rate test

Accelerated substitution in Clupeocephala CYP3B, C and D subfamilies was assessed by conducting relative rate tests between sequence pairs and between subfamily comparisons. Using chick CYP3A37 as outgroup, the differences in protein substitution rate along lineages were estimated in all combinations of two CYP3 subfamilies using RRTree (Robinson-Rechavi and Huchon 2000). Likelihood ratio test (LTR) of different rate of substitution was conducted between all possible Clupeocephala CYP3 pairwise comparisons with chicken CYP3A37 as outgroup using hypothesis testing using Phylogenies (HYPHY) (Pond et al. 2005). Null model assuming fixed branch length across lineages was compared to alternative model with unconstrained branch length along each lineage. LRT was conducted using chi-square test with 1 degree of freedom.

2.4 Functional divergence detection

To investigate adaptive functional diversification at amino acid level, detection of site-specific rate shift (type I functional divergence) among genes was performed using a two-state model (Gu 1999) implemented in the DIVERGE program (Gu and Vander Velden 2002). Type I function divergence describes altered evolutionary rate among gene clusters, thus indicating different functional constraint after gene duplication or speciation (Gu 1999). This type of functional divergence is measured by the coefficient of functional divergence θ, representing the decrease in the correlation of substitution rates in two gene clusters. The higher the θ value, the more pronounced the functional divergence between two gene clusters compared. A null model ($\theta = 0$) assumes no evolutionary rate difference between two gene clusters. Rejection of the null hypothesis provides statistical evidence for shifted site-specific rates between the gene clusters investigated (Gu 1999). LRT was conducted using chi-square test with 1 degree of freedom. The responsible amino acid sites were identified based on the posterior probability that any site has underwent functional divergence using a site-specific profile (Gu 1999). We examined the sequence data for phylogenetic tree construction. Five clusters (Primate, Glires and Laurasiatheria CYP3A, Clupeocephala CYP3A/D and Clupeocephala CYP3B/C) were defined (see Appendix Fig. 8.1 for details) so that each has both adequate number of sequences and sequence divergence. The primate CYP3A cluster comprises all available primate CYP3A sequences listed in Fig. 10. The input tree was taken from the Bayesian analysis.

2.5 Gene conversion

GENECONV program (Sawyer 1989) was used to detect gene conversion events among *CYP3* genes. *CYP3* coding sequences from each species or subfamilies were grouped to form separate datasets. To avoid effects of strong selection and recent mutation, only silent-site polymorphisms were used. The program was run with random-number seed set at 123 for reproducibility. p values were computed from 10,000 permutations. The analysis was performed with mismatch not allowed, or with penalty of 1, 2 and 3. Global p values lower than 0.05 are considered as evidence for gene conversion. For primate *CYP3A*, additional analysis was also performed on Multi-LAGAN (Brudno et al. 2003) aligned *CYP3A* genomic sequences and promoter sequences (Appendix Table. 8.1). The predicted gene conversion events involving short fragments with not enough informative sites were discarded. The redundant events and false-positive recombination events between genes from different species were filtered out.

2.6 Likelihood ratio tests for positive selection

Evolution of primate *CYP3A* sequences was assessed using the ratio of non-synonymous (amino acid altering) and synonymous (silent) substitution rates (= dn/ds = Ka/Ks = ω) as a measure. As a very conservative estimate, positive (Darwinian) selection ($\omega > 1$) is assumed if the non-synonymous substitution rate exceeds the synonymous substitution rate, due to an overall beneficial effect of amino acid exchanges on individual fitness. An excess of the synonymous substitution rate ($\omega < 1$) indicates negative selection of nonsynonymous substitutions, due to their detrimental effect on fitness. Neutral evolution ($\omega = 1$) indicates that non-synonymous substitutions do not affect individual fitness. The analyzed dataset comprised 18 genes from those primate species for which the complete *CYP3A* gene dataset was available (human, chimpanzee, orangutan, and rhesus). The analysis was performed using the maximum likelihood approach implemented in PAML package version 3.15 (Nielsen and Yang 1998; Yang 1998; Yang et al. 2000). The intree used was taken from above phylogenetic analyses. Site- and branch-specific analyses were carried out with Codeml. Sites with ambiguity data were not removed from the dataset (cleandata = 0). The codon frequency was estimated from a 3X4 matrix. We performed two parallel likelihood ratio tests (LRTs) for the presence of codon sites with $\omega>1$ (LRT I + II). A third LRT tested for lineage-specificity of ω (LRT III). For each of the LRTs, twice the log-likelihood difference between the alternative and the null model was compared to critical values from a chi-square distribution with degrees of freedom equal to the difference in the number of free parameters between both models (Table 1). LRT I and II compared discrete model M3 and beta&ω model M8 with the corresponding null models, i.e. one ratio model M0 and beta null model M7, respectively. To avoid local optima, M8 was run twice with initial ω values smaller or larger than one. Candidate sites for positive selection were pinpointed using the naive empirical Bayes approach (M3 + M8) and the Bayes empirical Bayes approach (M8). Only sites with a minimum support from posterior probability of 0.95 were further considered. LRT III compared free ratio model and one ratio model. For details of the implementation of each model, see references 26, 27, and 28. To pinpoint candidate sites of positive selection along branches of particular interest (Ho7 + Hs4, see Fig. 15), we inferred ancestral sequences using Baseml (model = HKY85, cleandata = 0). Subsequently, we identified nonsynonymous exchanges by pairwise comparison of the sequences at the ends of Ho7 and Hs4, respectively.

2.7 Analysis of primate *CYP3A* promoters

The available promoter sequences for *CYP3A* genes from human, chimpanzee and rhesus were retrieved from the UCSC genome browser. Baboon, marmoset and galago *CYP3A* promoter sequences were obtained from NCBI or assembled by BAC sequencing (Appendix Table8.1). Alltogether twenty promoter sequences were divided into three groups (twelve *CYP3A7/67/4/21/91/92*, five *CYP3A5* and three *CYP3A43*, see Fig. 20) based on their sequence similarity. Each group of sequences were aligned separately using Multi-LAGAN (Brudno et al. 2003). The potential nuclear receptor response elements (RE) in each promoter sequence were detected using Nuclear Hormone Receptor (NHR)-Scan (Sandelin and Wasserman 2005) with default settings. As numerous gaps were introduced into each sequence during the alignment procedure, the original position of each predicted RE was converted to its corresponding column number in the alignment using Bioperl module (Stajich et al. 2002) (www.bioperl.org), so that REs across different promoters could be compared. Repeat elements within the promoter sequences were identified using RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). The RE and repeat element annotations were parsed using custom Perl scripts and all predicted REs were mapped to a matrix as shown (Figs 20 and 21). The REs belonging to same category and with identical column number in different promoters were defined as homologs.

2.8 Tissue and RNA samples

Chimpanzee, orangutan and rhesus liver samples were obtained from the Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany, olive baboon and hamadryas baboon (*Papio anubis*) liver samples from Deutsches Primatenzentrum Göttingen (DPZ). Total RNA from the liver of additional rhesus individuals was purchased from BioChain Institute, Inc (Hayward, CA, USA). Marmoset cDNA was obtained from University of Gottingen, Germany. For tissue samples, the RNA was extracted using TRIzol following a standard protocol. Briefly, tissue samples were homogenized (50-100 mg with 1 ml of TRIzol) and incubated at room temperature for 10 minutes. 1/5 column of chloroform was added into the homogenate and was shaken heavily by hand for 15 seconds. After 10 minutes’ incubation at room temperature, the samples were centrifuged at 13,000 g for 15 minutes at 4 °C. The aqueous phase was transferred into a new tube for RNA isolation and the remaining interphase and organic phase were used for genomic DNA preparation. RNA was precipitated from the aqueous phase by adding an equal volume of isopropylalcohol. After brief vortexing, the probes were incubated at room temperature for 10 minutes and then centrifuged at 10,000 g for 10 minutes. The pellets were washed twice with 75% ethanol, air dried, and dissolved in RNase-free water. For DNA isolation, ethanol (1/3 volume of the original TRIzol volume) was added to the combined interpase and organic phases. The probes were gently vortexed, incubated at room temperature for 5 minutes, and centrifuged at maximal speed for 10 minutes. The pellets were washed twice in 0.1 M sodium citrate for 30 minutes and dissolved in sterile water.

2.9 Cloning of primate *CYP3A* coding regions

One µg of total RNA was subjected to reverse transcription in a final volume of 20 µl using SuperScript transcriptase (Invitrogen, Germany), following the manufacturer's instructions. cDNA derived from approximately 100 ng total RNA was subjected to PCR amplification. All reactions were performed in a 25-µl reaction mixture containing 200 pmol of each of the forward and reverse primers (Operon, Germany), 2 mM $MgCl_2$, 200 µM dNTPs, and 0.25 units of *Taq* DNA polymerase. The reactions were carried out for 35 cycles, with 10 s at 94 °C, 30 s at the proper annealing temperatures, and 2 min at 72 °C. Initial amplification was performed using primers from exons 1 and 13 of each gene. In case of low yield, the PCR product was diluted 100-fold and followed by nested PCR using internal primers. The primers used in the PCR are listed in the Appendix, Table 8.2. PCR products were purified with the Cycle Pure Kit (peQlab, Erlangen, Germany). In case of co-amplication of nonspecific products, completed PCR reactions were separated on ethidium bromide stained agarose gel in 1 X TAE buffer and visualized by UV-illumination. DNA bands with the expected size were recovered using Gel Extraction Kit (peQlab, Erlangen, Germany). The products of RT-PCR reactions were sequenced, either directly or following AT cloning, by GENterpise (Mainz, Germany).

2.10 Screening for potential *CYP3A67* in humans

To test if loss of *CYP3A67* was fixed in the human lineage, part of 3'UTR regions of *CYP3A67* and *CYP3A7* were co-amplified with the primer pair Alu-F (5'-AATGGGCAAAGTCATAGTG-3') and Alu-R (5'-GCTTCTCCTAGGACTATCTTCA-3'). Due to the high sequence similarity between *CYP3A67* and *CYP3A7*, primers that specifically amplify *CYP3A67* were not sought. The co-amplified 3'UTR region in *CYP3A7* differs from that of *CYP3A67* in that it contains an extra AluY insertion of ~250 bp. Thus, the smaller fragment size (393 bp) indicates the possible existence of *CYP3A67* in the sample, whereas the larger fragment (642 bp) indicates the presence of *CYP3A7*. 99 Central Europeans and 38 African (Bantu) DNA samples were subjected to PCR diagnostics. Chimpanzee BAC DNA (CH251-400N23) encompassing the 3'UTR of both *CYP3A7* and *CYP3A67* was used as a positive control for PCR reaction. The PCR reactions were carried out in a total volume of 25 µl containing 5~10 ng of genomic DNA at an annealing temperature of 54 °C. Other parameters of the PCR program were the same as described for cloning of primate *CYP3A* coding regions.

2.11 Diagnostics of the *CYP3A67* polymorphism in chimpanzees

A comparison of the two BAC clones sequences (CH251-35M15 and CH251-171D20) together with the HTG sequence from NCBI (Genbank accession No. AC145951.2) indicated that a non-canonical *CYP3A67* allele may be present in chimpanzees. This allele contains a deletion encompassing region form the proximal promoter of *CYP3A67* to the first exon of the upstream pseudogene (Fig. 4).

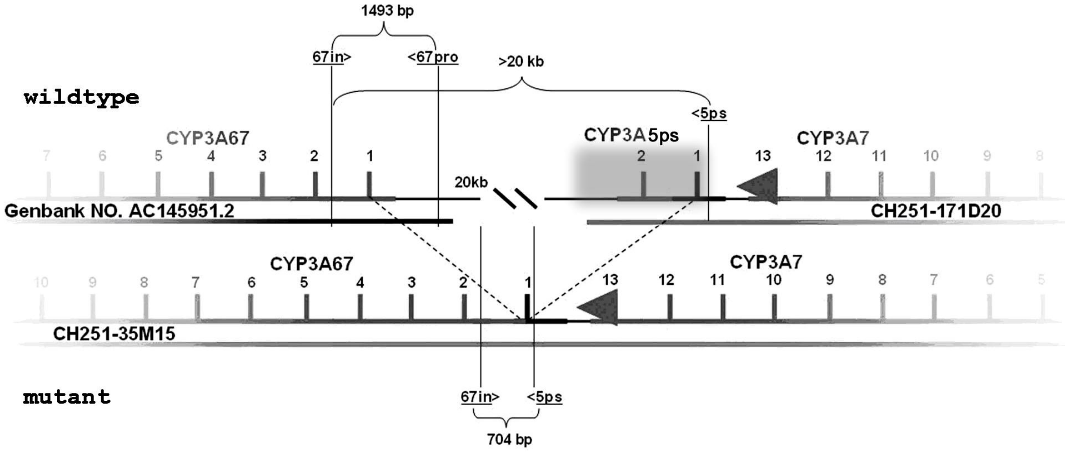


Fig.4. Strategy for PCR diagnostics of two putative *CYP3A67* alleles. Gene structures are shown in detail with identity of each exon labeled by number (e.g., 13: exon13). The *CYP3A5* pseudogene (*CYP3A5ps*) is shadowed. Targeting sites, extension direction and expected size of products of each primer pairs are shown.

To confirm the existence of the two different *CYP3A67* alleles in chimpanzees, nested-PCR was carried out with different primer pairs which distinguish between them. Forward primers 67in (5'-CTGCAAAACATCCACCATAACTTC-3') and 67in2 (5'-ACCACCATGCCCAGGTAACT-3') match parts of *CYP3A67* intron 1 which are present in both alleles. Reverse primer 67pro (5'-CAGAGGATCAGCCTGAAAATGC-3') specifically binds to *CYP3A67* promoter but not to the *CYP3A7* promoter. Reverse primer 5ps (5'-CATAAACATTTTAGCAGCTTGACTTAAG-3') was designed against the proximal promoter of the *CYP3A5* pseudogene downstream of *CYP3A7*. DNA from 5 chimpanzee BAC clones and 4 different chimpanzee individuals were investigated. The first round PCR was performed with forward primer 67in and each of the reverse primers (67pro and 67ps). The PCR products were diluted and further amplified with forward primer 67in2 and each of the reverse primers. The PCR conditions were similar to the above mentioned, with an annealing temperature of 56 °C. PCR products obtained with the primer pair 67in/67in2 and 5ps indicate the presence of the allele which contains the *CYP3A67* promoter deletion, whereas PCR products obtained with the primer pair 67in/67in2 and 67pro represent the wildtype *CYP3A67* allele (Fig. 4).

3. Results

3.1 Phylogenomics of *CYP3* loci

A total of 25 loci (Fig. 5 and Appendix Table 8.3) containing full-length *CYP3* genes were identified in the 16 genomes queried by Blast with representative *CYP3* sequences from each subclade. The number of *CYP3* genes in the investigated species ranges from 2 (chicken) to 10 (mouse). *CYP3B*, *C*, and *D* genes were found only in Clupeocephala. In most vertebrates, apparently intact *CYP3* genes are located within one of two *CYP3* Homologous Regions, *CYPHR1* and *CYPHR2*. *CYP3HR1* (*SDK1-CYP3-FOXK1-KIAAO415-FLJ10324*) harbors *CYP3* genes in all non-Eutherian vertebrates but frog, tetraodon, and medaka. In Eutheria (placental mammals), *CYP3HR1* contains only *CYP3A* gene remnants (referred to as *CYP3A84P* in D. Nelson’s Cytochrome P450 Homepage), whereas all apparently intact *CYP3A* genes are found within the *CYP3HR2* (*CPFS4-ATP5J2-ZNFs-CYP3A-OR2AE1-TRIM4-GJE1-AZJP1*). In rodents (rat and mouse), the *CYP3HR2* got split into two parts by two independent genomic rearrangement events, leading to two *CYP3A* loci in each species separated by about 8 Mb. The *CYP3HR1* in some Clupeocephala species contains *CYP3C* or *CYP3D* genes, in stickleback together with *CYP3A* genes. Several additional *CYP3*-containing genomic regions were found outside *CYPHR1* and *CYPHR2* in the frog and in the Clupeocephala (Fig. 5). All members of the *CYP3B* subfamily are contained in a syntenic region (*KCNH-CYP3B*), whereas all frog *CYP3A* genes were found in three loci specific for amphibians.

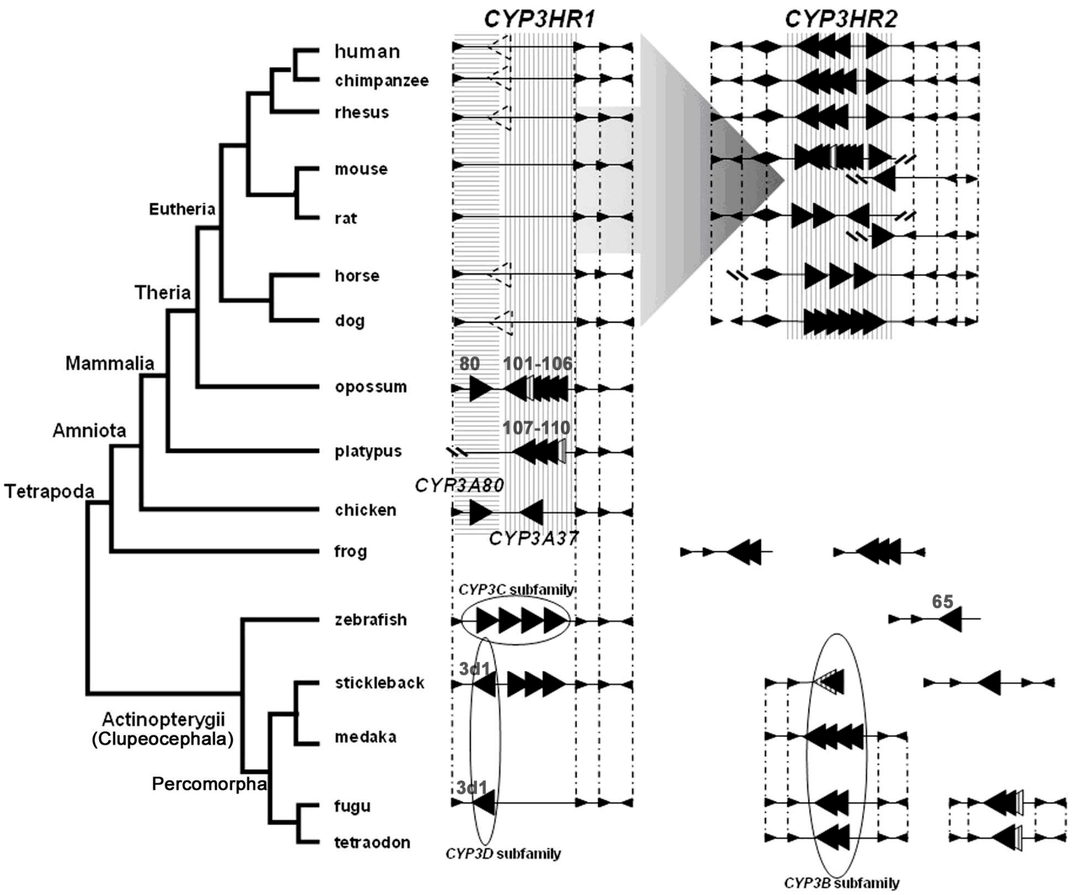
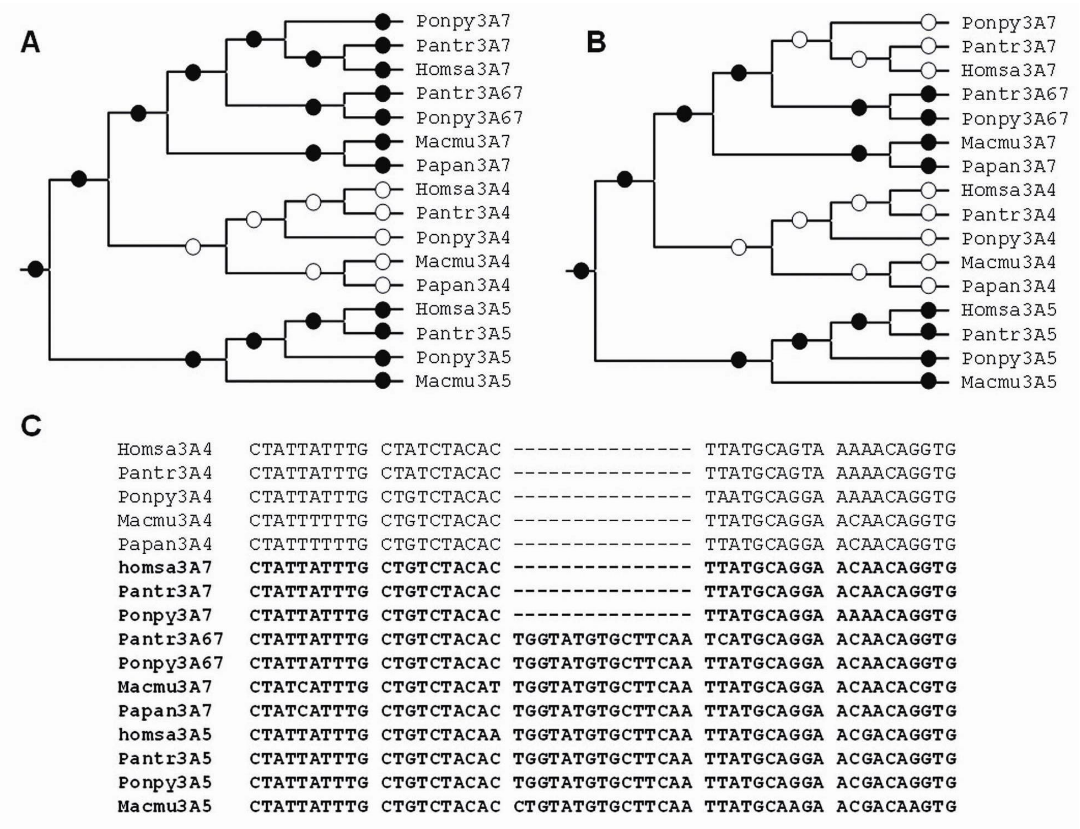


Fig. 5. Comparison of *CYP3* loci from 16 different species. Large triangles indicate *CYP3* genes and small triangles indicate flanking genes. Members of *CYP3B*, *CYP3C* and *CYP3D* subfamilies are circled for easier identification; all other large black triangles are *CYP3A* genes. Remnants of ancient *CYP3A* genes in mammalian *CYP3HR1* (Homology Region 1) are depicted by dashed line as large transparent triangles. Near-full length *CYP3* pseudogenes are represented by striped triangles. *CYP3A80*- and *CYP3A37*-related genes found in amniota species are shadowed by grey color in different patterns. Critical *CYP3A* genes are label with their ID numbers (e.g., 65: *CYP3A65*). The genomic translocation of *CYP3A* from *CYP3HR1* to *CYP3HR2* early in Eutheria evolution is indicated by a large gray arrow. See Appendix Table 8.3 for detail information and genomic coordinates of each locus.

3.2 Gene conversion in *CYP3* genes

A total of 9 potential gene pairs were identified which might have been affected by gene conversion events involving the protein-coding regions. These included *CYP3* from rat, dog, pig, medaka, and some primate species (Appendix Table 8.4). One of the detected primate gene conversion events replaced exon 6 and a part of intron 6 of *CYP3A7* by the corresponding portion of *CYP3A4* in a common hominidae ancestor, since it is detectable in human, chimpanzee, and orangutan sequences. This is also supported by a deletion found in *CYP3A4* and *CYP3A7* genes in these 3 species, but absent from these genes in rhesus and baboon, and from *CYP3A67* and *CYP3A5* genes in all species (Fig. 6). Moreover, analysis of full length genomic sequences revealed a number of additional gene conversion events restricted to intron sequences (data not shown).



```
Homsa3A4    CTATTATTTG CTATCTACAC --------------- TTATGCAGTA AAAACAGGTG
Pantr3A4    CTATTATTTG CTATCTACAC --------------- TTATGCAGTA AAAACAGGTG
Ponpy3A4    CTATTATTTG CTGTCTACAC --------------- TAATGCAGGA AAAACAGGTG
Macmu3A4    CTATTTTTTG CTGTCTACAC --------------- TTATGCAGGA ACAACAGGTG
Papan3A4    CTATTTTTTG CTGTCTACAC --------------- TTATGCAGGA ACAACAGGTG
homsa3A7    CTATTATTTG CTGTCTACAC --------------- TTATGCAGGA ACAACAGGTG
Pantr3A7    CTATTATTTG CTGTCTACAC --------------- TTATGCAGGA ACAACAGGTG
Ponpy3A7    CTATTATTTG CTGTCTACAC --------------- TTATGCAGGA AAAACAGGTG
Pantr3A67   CTATTATTTG CTGTCTACAC TGGTATGTGCTTCAA TCATGCAGGA ACAACAGGTG
Ponpy3A67   CTATTATTTG CTGTCTACAC TGGTATGTGCTTCAA TTATGCAGGA ACAACAGGTG
Macmu3A7    CTATCATTTG CTGTCTACAT TGGTATGTGCTTCAA TTATGCAGGA ACAACACGTG
Papan3A7    CTATCATTTG CTGTCTACAC TGGTATGTGCTTCAA TTATGCAGGA ACAACAGGTG
homsa3A5    CTATTATTTG CTGTCTACAA TGGTATGTGCTTCAA TTATGCAGGA ACGACAGGTG
Pantr3A5    CTATTATTTG CTGTCTACAC TGGTATGTGCTTCAA TTATGCAGGA ACGACAGGTG
Ponpy3A5    CTATTATTTG CTGTCTACAC TGGTATGTGCTTCAA TTATGCAGGA ACGACAGGTG
Macmu3A5    CTATTATTTG CTGTCTACAC CTGTATGTGCTTCAA TTATGCAAGA ACGACAAGTG
```

Fig. 6. The distribution of the intron 6 15-bp-indel among primate *CYP3A* phylogeny, before (A) and after (B) the gene conversion event between hominid *CYP3A4* and *CYP3A7*. The absence and presence of the 15-bp-indel is indicated by white and black circles, respectively. (C) The alignment of the indel and the flanking sequences from the *CYP3A* genes studied. Homsa (*Homo sapiens*), Pantr (*Pan troglodytes*), Macmu (*Macaca mulatta*), Ponpy (*Pongo pygmaeus*), Papan (*Papio anubis*).

3.3 Phylogeny of vertebrate *CYP3*

A schematic representation of a phylogenetic tree based on 107 CYP3 protein coding sequences is given in Fig. 7. The entire tree is provided in the Appendix (Fig. 8.1). Phylogenetic analyses using Bayesian and maximum likelihood methods produced similar trees, the only difference being the relationships of rodents, primates and Laurasiatheria CYP3A within the well-supported eutherian *CYP3A* clade. Although the posterior probability and the bootstrap support values are low (Appendix, Fig. 8.1), all extant amnioid *CYP3A* form two distinct groups (Fig. 7 and Appendix Fig. 8.1). The first group ("*CYP3A80* clade" in Fig. 7) includes bird *CYP3A80*, three anolis *CYP3A* genes, all frog *CYP3A*, and one *CYP3A* gene from opossum. The second group ("*CYP3A37* clade" in Fig. 7) comprises bird *CYP3A37*, one anolis *CYP3A*, most *CYP3A* genes from opossum and platypus, and all eutherian *CYP3A* genes. This data together indicates that all extant Amniota *CYP3A* genes are derived from two *CYP3A* genes (ancestors of *CYP3A37* and *CYP3A80*) in early Amniota. The absence of extant eutherian *CYP3A* in the *CYP3A80* clade suggests that *CYP3A80* orthologs were lost early in Eutheria evolution. Therefore, all *CYP3A* genes in Eutheria are derived from an ancient *CYP3A37* ortholog.

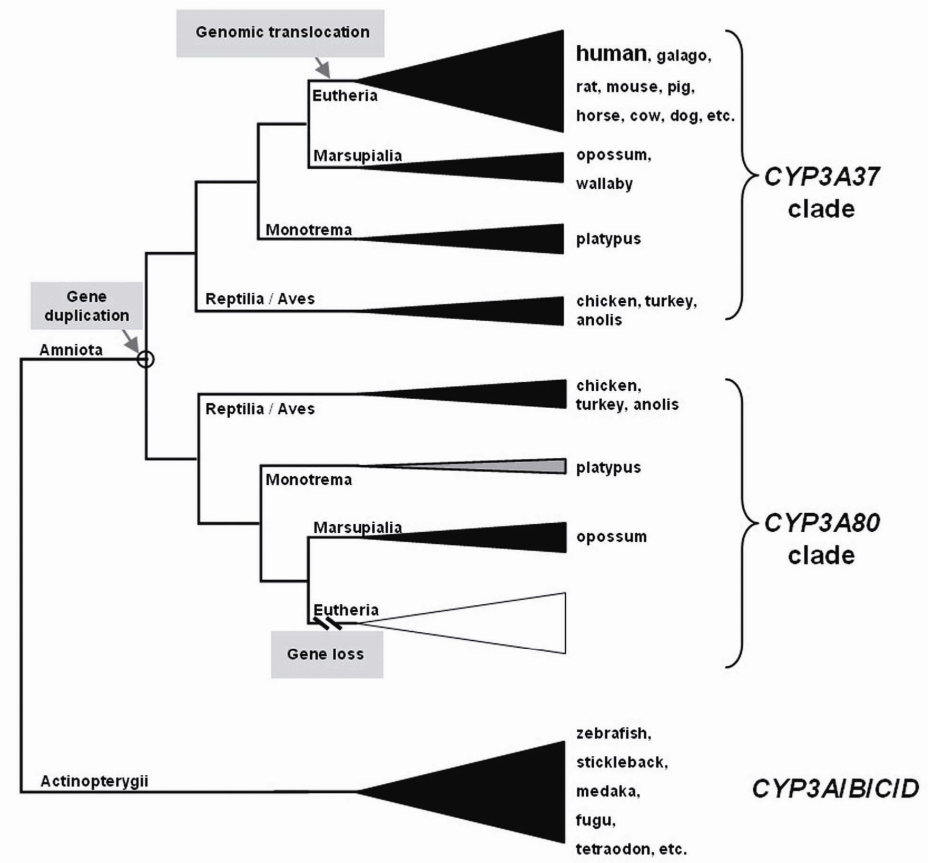


Fig. 7. Schematic representation of the *CYP3* phylogeny based on 107 protein sequences from 31 species. Each triangle represents a *CYP3* cluster. The triangle in grey color indicates incomplete sequences not included in tree construction. *CYP3A80*-related genes expected, but not found in Eutheria, are indicated by a white triangle. The animals on the right of each triangle represent the taxa from which *CYP3* sequences were used for phylogenetic analysis. For detailed phylogenetic tree, see Appendix Fig. 8.1. The major events in the Amniota ancestor (gene duplication resulting in ancestors of *CYP3A37* and *CYP3A80*), and in the eutherian ancestor (genomic translocation of the *CYP3A37* ortholog from *CYP3HR1* to *CYP3HR2* and the loss of *CYP3A80* ancestor), are indicated.

3.4 Relative rate tests

Relative rate tests among groups resulted to extremely low p values in all comparisons between CYP3A and non-CYP3A subclades. Of the remaining tests, only the comparison between CYP3B and CYP3C is statistically significant without Bonferroni correction. The substitution rate in CYP3B, C and D is higher than that of CYP3A, whereas CYP3B evolved at rate higher than CYP3C (Table 2). Consistent with the comparison between groups, log likelihood ratio larger than the critical value 3.85 (cutoff at 5%) was found for nearly all (255/257) pairwise comparisons between CYP3A and non-CYP3A proteins, and for most (34/36) pairwise comparisons between CYP3B and CYP3C proteins. A number of significant p values were also found for tests involving CYP3D, either within the group, or in comparison with CYP3C and CYP3B (Table 3).

Table 2. Statistics of relative rate test between groups

Group1/Group2	K1	K2	dK	sd_dK	Ratio	p value
CYP3A/CYP3D	0.54	0.72	-0.18	0.037	-4.74	2.79E-06**
CYP3A/CYP3C	0.54	0.68	-0.14	0.036	-3.74	0.000186**
CYP3A/CYP3B	0.54	0.77	-0.22	0.038	-5.85	1.00E-07**
CYP3D/CYP3C	0.72	0.68	0.041	0.048	0.85	0.394421
CYP3D/CYP3B	0.72	0.77	-0.05	0.047	-0.98	0.325594
CYP3C/CYP3B	0.68	0.77	-0.09	0.038	-2.26	0.023518*

K1 (K2): the mean values of protein distances between the proteins in group 1 (or group 2) and outgroup. dK: difference of K1 between K2. sd: standard deviation. Ratio: the dK-to-sd_dK ratio. p value: uncorrected p value for each test. * Significant tests with before Bonferroni correction. ** Significant tests with after Bonferroni correction.

Table 3. Number of relative rate tests according to p values for pairwise comparisons

Critical value*	2.71	3.85	6.64	7.87	10.82	>11
p value*	< 0.1	0.05~0.1	0.01~0.05	0.005~0.01	0.001~0.005	< 0.001
CYP3A/CYP3A	111	6	2	0	0	0
CYP3B/CYP3B	34	2	0	0	0	0
CYP3C/CYP3C	6	0	0	0	0	0
CYP3D/CYP3D	1	0	1	1	0	0
CYP3A/CYP3B	0	0	1	0	0	**144**
CYP3A/CYP3C	0	2	21	18	12	**11**
CYP3A/CYP3D	0	0	3	1	4	**40**
CYP3B/CYP3C	0	2	8	6	12	**8**
CYP3B/CYP3D	23	5	7	2	0	0
CYP3C/CYP3D	8	1	2	1	0	0

* Critical values and their corresponding p values from chi-square distribution with one degree of freedom. Numbers of significant pairwise tests at 5% cutoff are shadowed.

3.5 Evidence for functional divergence

The functional coefficients (θ) between gene clusters ranged from about 0.1 to 0.47 (Table. 4). The highest θ was observed for tests comparing Clupeocephala and mammalian CYP3 clusters, followed by the Clupeocephala CYP3A/D and CYP3B/C comparisons. All tests among mammalian CYP3A comparisons resulted in low values of θ. Except for the two tests (Laurasiatheria vs. Primate and Primate vs. Glires), all the other tests rejected the null hypothesis (θ = 0) with strong statistical support after multiple test correction.

Table 4. The coefficients of functional divergence between pairwise of CYP3 clusters

	Clu 3AD / Clu 3BC	Clu 3AD / Glires 3A	Clu 3AD / Primate 3A	Clu 3AD / Laurasiatheria 3A	Clu3BC / Laurasiatheria 3A
Theta (θ)	**0.2664**	**0.3144**	**0.3864**	0.3728	**0.4232**
SE Theta	0.048204	0.047703	0.050779	0.047547	0.046516
LRT	30.54259*	43.438741*	57.904436*	61.475965*	82.7728*

	Clu3BC/ Primate 3A	Clu 3BC/ Glires 3A	Laurasiatheria 3A/ Primate 3A	Laurasiatheria 3A/ Glires 3A	Primate 3A / Glires 3A
Theta (θ)	**0.468**	**0.3856**	0.096	0.18	0.124
SE Theta	0.051743	0.048731	0.03854	0.039414	0.04776
LRT	81.807576*	62.611607*	6.204803	20.856931*	6.740978

The significant tests at 5% cutoff after multiple test correction are labeled with *. CYP3A, B, C and D were shortened as 3A, 3B, 3C and 3D, respectively. Clupeocephala CYP3A/D and CYP3B/C were shortened as Clu 3AD and Clu 3BC, respectively.

The posterior probility (>= 0.5) of any site to be functionally divergent was plotted against the amino acid position (Fig. 8). The relevant sites were distributed throughout the protein. The number of functionally divergent relevant sites is generally correlated with the value of functional coefficients from each test. Highest number of sites was found in the mammalian CYP3A vs. Clupeocephala CYP3B/C comparison followed by the mammalian CYP3A vs. Clupeocephala CYP3A/D comparison. Only few sites are predicted to be functionally relevant when mammalian CYP3A clusters were compared to each other. Detailed examples for amino acid positions where functional divergence is likely to have occurred are shown in Table 5.

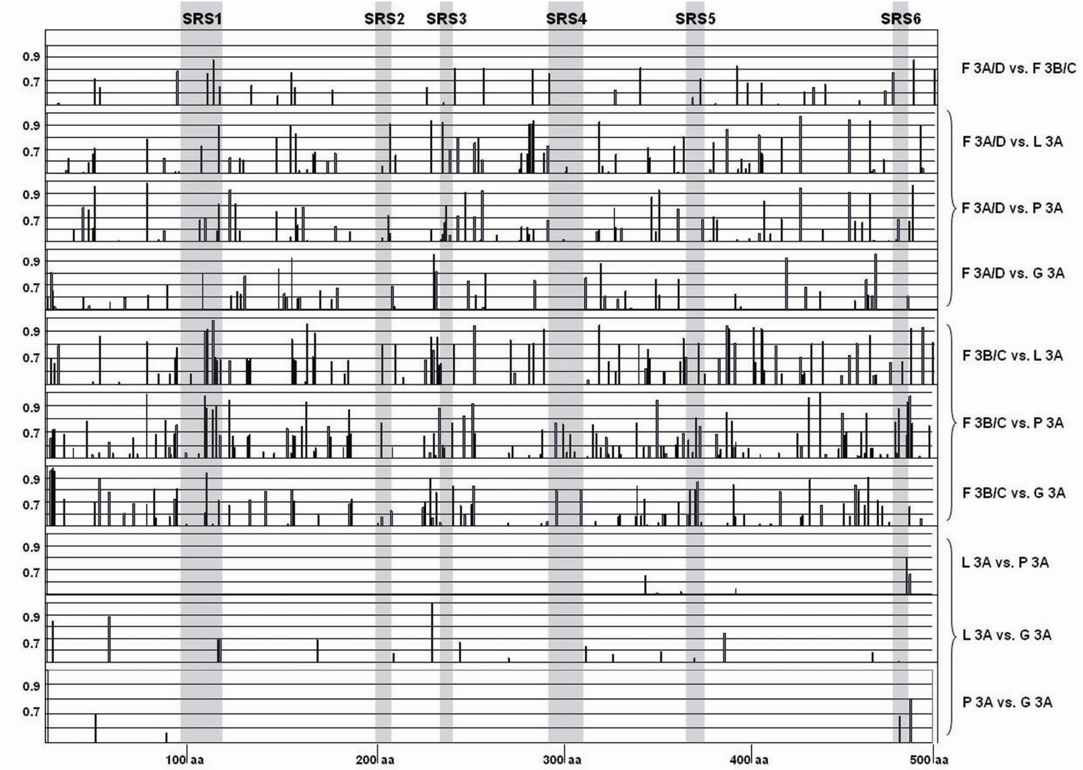


Fig. 8. The plot of posterior probability for each CYP3 amino acid site having undergone functional divergence. X axis: position along CYP3A protein. Y axis: the posterior probility of being involved in functional divergence. The regions of six Substrate Recognition Sites (SRS1-6) (Gotoh 1992) are shadowed. F 3A/D: Clupeocephala CYP3A/D, F 3B/C: Clupeocephala CYP3B/C, P 3A: Primate CYP3A cluster, G 3A: Glires CYP3A, L 3A: Laurasiatheria CYP3A.

Table 5. Examples of amino acid sites which underwent functional divergence

	Conserved cluster	Altered cluster	Position
clusters	Laurasiatheria 3A	Clupeocephala 3B/C	
amino acid	GRGGGGGGGGGGGGGGG	DDGDLGAFIEPDI	109
clusters	Clupeocephala 3B/C	Glires 3A	
amino acid	WWWWWWWWWWWWWW	SHHSRHRRRRRYHYYYYYYY	28
clusters	Clupeocephala 3A/D	Primate 3A	
amino acid	RRRRRRRRRRRRRRRRRRR	RRRRQRRLLQQQLLRRQQQQQQQQQQQR	78
clusters	Clupeocephala 3A/D	Primate 3A	
amino acid	PPPPPPPPPPPPPPPPPPP	TTTTPPTTPTTTTTTTTSSPPPAAPPPPPP	485
clusters	Glires 3A	Laurasiatheria 3A	
amino acid	TTTTTTTTTTTTTITTTTTTTTT	NQRKNITTTTTTTSVVVIN	230
clusters	Glires 3A	Clupeocephala 3A/D	
amino acid	NNNNNNNNNNNNNNNNNNNNN	SSQQQTRTRRRRRNNKHHHN	462

"CYP3 cluster" rows show the two CYP3A clusters that were tested for functional divergence (shifted site-specific rates). "amino acid" rows provide the amino acid at a given position in all sequences from the two protein clusters tested. Position: refers to the corresponding positions in the human CYP3A4 protein sequence.

3.6 Genomics and phylogenetics of primate *CYP3A*

Four and five *CYP3A* genes have been reported, respectively, within the human (Gellner et al. 2001) and chimpanzee *CYP3A* loci (Williams et al. 2004a). We analyzed *CYP3A* loci in three more primate species representing key evolutionary positions in the primate phylogeny: rhesus (Old World monkey), common marmoset (New World monkey), and galago (Strepsirrhini). The rhesus *CYP3A* locus resides on chromosome 3. It contains four complete genes (*CYP3A5*, *7*, *4* and *43*), and several pseudogenes composed of detritus exons (Nelson et al. 2004) in an organization resembling that found in the human locus (Fig. 9, See Appendix Fig. 8.2 for details). A ~40 kb genomic insertion consisting mainly of low copy number repeats was found between rhesus *CYP3A7* and *CYP3A4*. Sequence analysis of a baboon *CYP3A*-containing BAC (Genbank No, AC141417.16) revealed a similar insertion between baboon *CYP3A7* and *CYP3A4* genes. Note that the previously released two rhesus *CYP3A* sequences, designated as *CYP3A64* and *CYP3A66* are orthologs of human *CYP3A4* and *CYP3A5* in rhesus. Thus, these two sequences are referred to as rhesus *CYP3A4* and *CYP3A5* to underscore the evolutionary relationships with other primate *CYP3A*. The marmoset *CYP3A* locus (Genbank accession No. EF600757 and EF600758) contains three genes (*CYP3A5*, *90* and *21*) in tandem (Fig. 9, See Appendix, Fig. 8.3 for details). No detritus exons were found in the *CYP3A* intergenic regions. In addition, a processed pseudogene (as evidenced by lack of exon 1 and of all introns, and by multiple in-frame stop codons) related to *CYP3A21* and of unknown genomic location, was found by assembling marmoset WGS available through the Trace Archive (Appendix 8.6). Galago *CYP3A* locus (Genbank accession No. EF600755 and EF600756) contains at least three tandem *CYP3A* genes of which one is in the opposite orientation to the others (Fig. 9, See Appendix Fig. 8.4 for details).

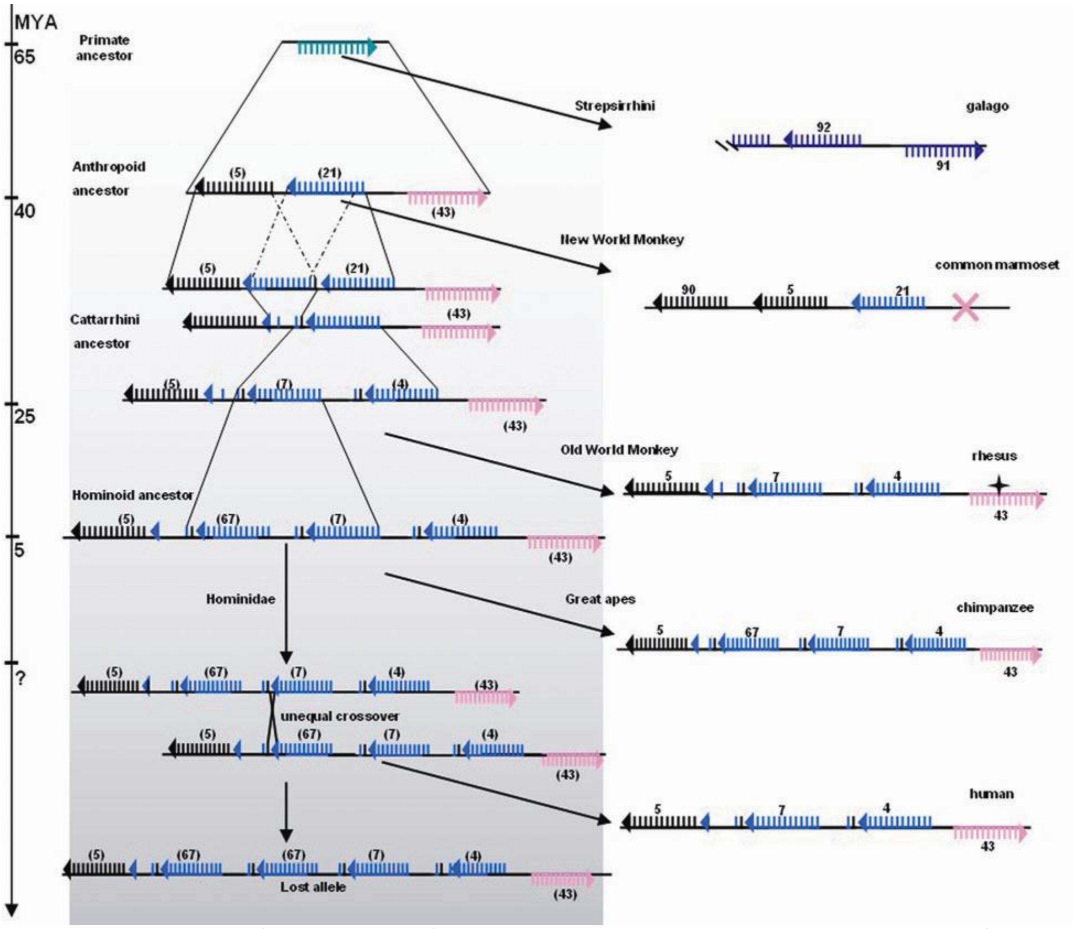


Fig. 9. A reconstruction of the evolution of primate *CYP3A* loci. Triangles represent exon 13 in functional *CYP3A* genes or in pseudogenes. The X represents a gene loss in the common marmoset. The asterisk represents pseudogenization of rhesus *CYP3A43*. Numbers indicate the gene identity (e.g. 5 = *CYP3A5*). The left part of the figure depicts the ancestral loci, including the apparently lost *5-67-67-7-4-43* allele resulting from the unequal crossover in the human lineage. *CYP3A* loci in selected contemporary primates are shown on the right. Time points of divergence are shown on the y-axis on the left (MYA, million years ago)

We also examined orangutan *CYP3A*-related WGS data in the NCBI Trace Archive. Although the entire structure of orangutan *CYP3A* locus was not obtained, we found the orthologs of all five chimpanzee *CYP3A* genes in this species. All genomic sequence-based *in silico* gene predictions of chimpanzee, rhesus, orangutan and common marmoset *CYP3A* genes were either supported by mRNA sequences in Genbank, or confirmed using RT-PCR of liver RNA followed by cloning and sequencing (Genbank accession No. EF589790~EF589808). The exceptions were the chimpanzee and orangutan *CYP3A7* genes which, similarly to human *CYP3A7*, might be expressed predominantly prenatally, and galago, liver samples of which were not available. In addition, spliced variants were identified for chimpanzee and rhesus *CYP3A5* (Genbank accession No. EU636002 and EU636003). A chimerical transcript was also amplified which is composed of rhesus *CYP3A4* exon 1-3 and *CYP3A43* exon 13 (Genbank accession No. EU636004). Since its splicing junctions all apply to the GT/AG rule, this transcript likely represents a true trans-splicing event rather than an artifact. The analysis revealed that gaps within the *CYP3A* locus in the first release of the chimpanzee genome sequence (PanTro1) led to a mix-up between the first exon of *CYP3A67* and the first detritus exon of a *CYP3A* pseudogene in previous publications (Rodriguez-Antona et al. 2005; Williams et al. 2004a).

The reconstructed process of primate *CYP3A* locus evolution shown in Fig. 9 is in agreement with primate *CYP3A* phylogeny (Fig. 10). The galago *CYP3A* resides on the basal position of the primate phylogenetic tree. Marmoset *CYP3A21* shares a common ancestor with catarrhine *CYP3A4*, *7* and *67*, whereas marmoset *CYP3A5* and *CYP3A90* share a common ancestry with catarrhine *CYP3A5*. Human *CYP3A*, *CYP3A4*, *5* and *43*, each have 1 to 1 orthologs in all Catarrhine species studied, whereas *CYP3A7* orthologs are restricted to Hominidae. No human *CYP3A* genes have 1 to 1 orthologs in the marmoset. The primate *CYP3A* phylogeny was strongly supported by orthologous insertion patterns of RGCs (10 retroposed elements and 44 indels) in various loci within primates *CYP3A* genes (Fig. 11). Only two conflicting RGCs were found, of which one is covered by the gene conversion shown in Fig. 6.

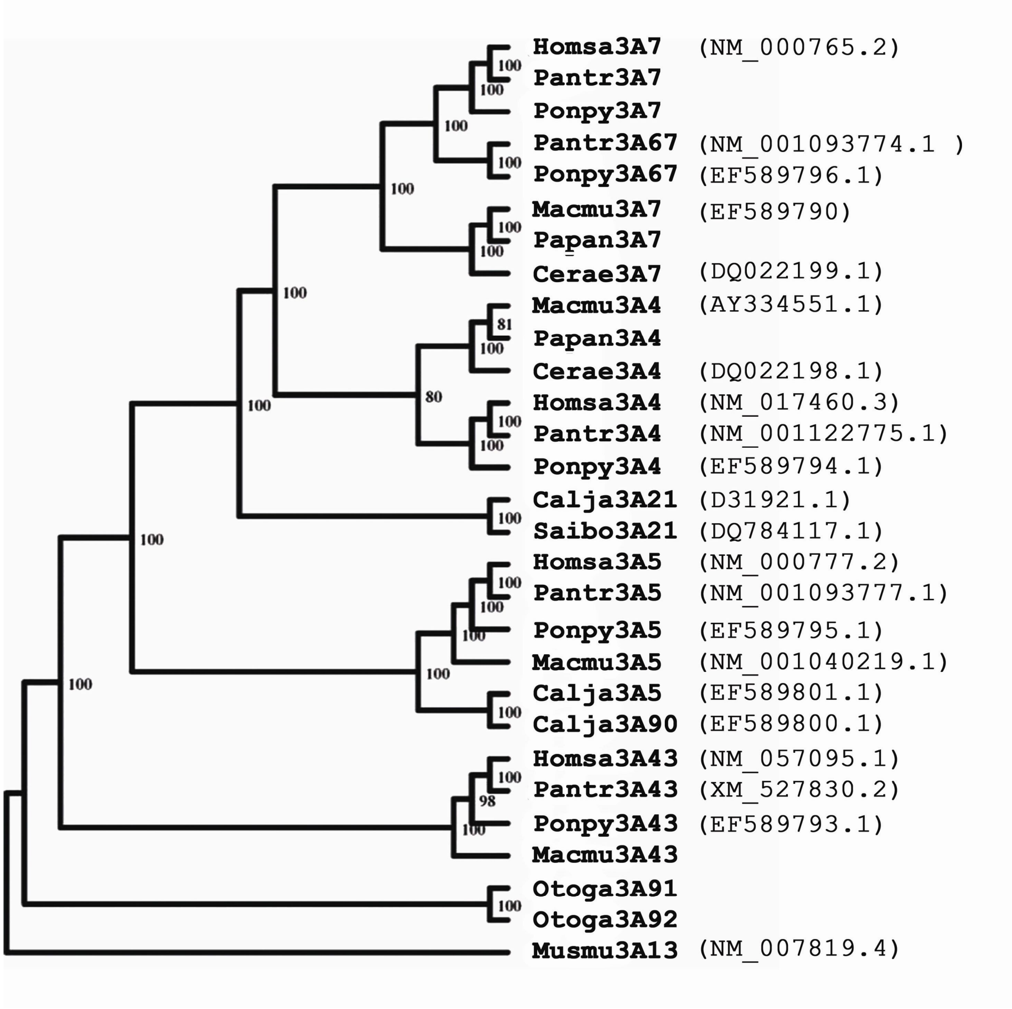
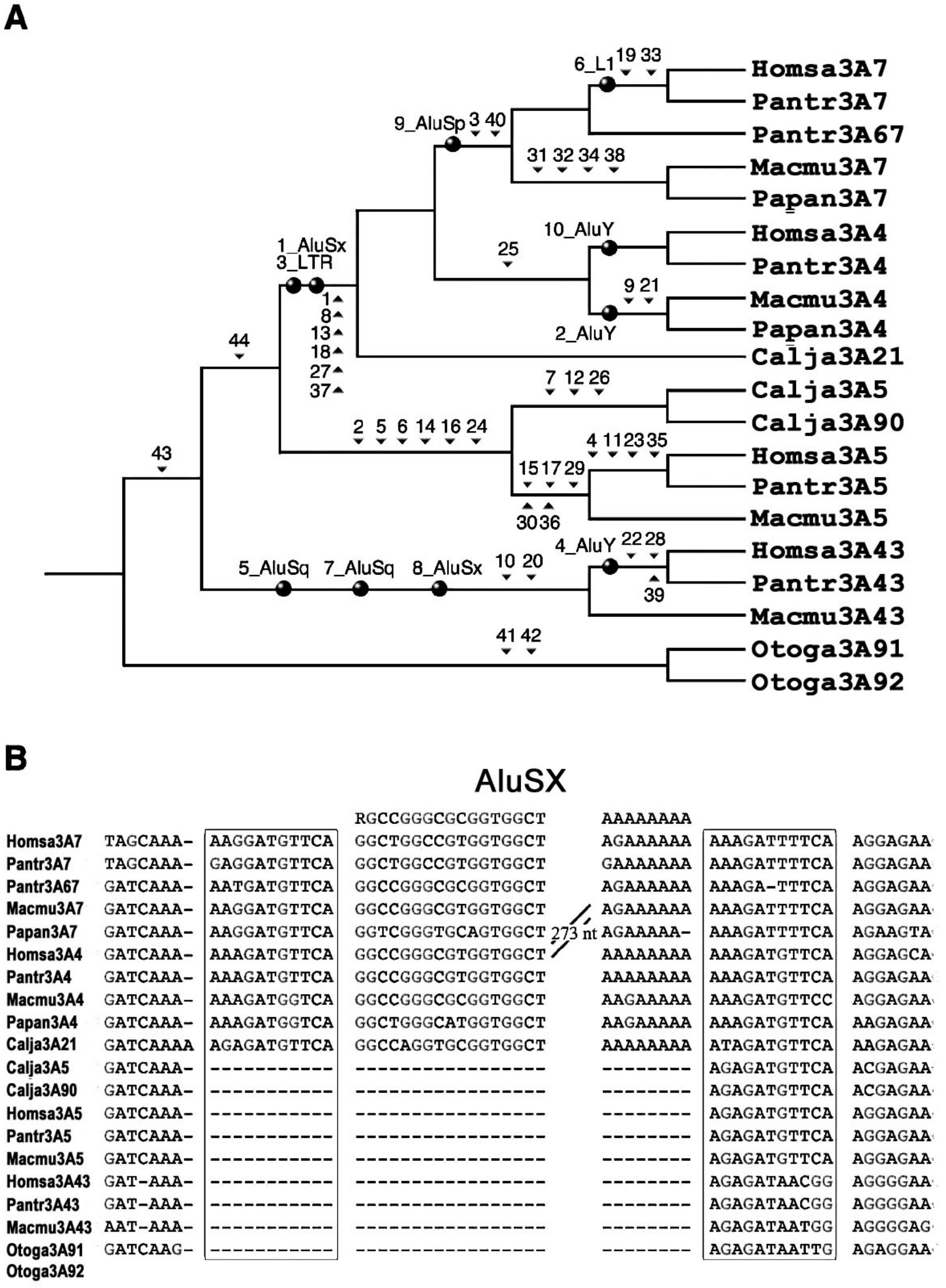


Fig. 10. The phylogenetic tree of primate *CYP3A* rooted by mouse *CYP3A13* based on the Bayesian analysis of full length *CYP3A* coding region. The posterior probabilities are given at each node. The sequence sources (Genbank accession number), if available, were provided in the parentheses. Cerae (*Cercopithecus aethiops*), Saibo (*Saimiri boliviensis*), Otoga (*Otolemur garnettii*), Mmsmu (*Mus musculus*).



			RGCCGGGCGCGGTGGCT		AAAAAAAA		
Homsa3A7	TAGCAAA-	AAGGATGTTCA	GGCTGGCCGTGGTGGCT		AGAAAAAA	AAAGATTTTCA	AGGAGAA
Pantr3A7	TAGCAAA-	GAGGATGTTCA	GGCTGGCCGTGGTGGCT		GAAAAAAA	AAAGATTTTCA	AGGAGAA
Pantr3A67	GATCAAA-	AATGATGTTCA	GGCCGGGCGCGGTGGCT		AGAAAAAA	AAAGA-TTTCA	AGGAGAA
Macmu3A7	GATCAAA-	AAGGATGTTCA	GGCCGGGCGTGGTGGCT		AGAAAAAA	AAAGATTTTCA	AGGAGAA
Papan3A7	GATCAAA-	AAGGATGTTCA	GGTCGGGTGCAGTGGCT	273 nt	AGAAAAA-	AAAGATTTTCA	AGAAGTA
Homsa3A4	GATCAAA-	AAAGATGTTCA	GGCCGGGCGTGGTGGCT		AAAAAAAA	AAAGATGTTCA	AGGAGCA
Pantr3A4	GATCAAA-	AAAGATGTTCA	GGCCGGGCGTGGTGGCT		AAAAAAAA	AAAGATGTTCA	AGGAGAA
Macmu3A4	GATCAAA-	AAAGATGGTCA	GGCCGGGCGCGGTGGCT		AAGAAAAA	AAAGATGTTCC	AGGAGAA
Papan3A4	GATCAAA-	AAAGATGGTCA	GGCTGGGCATGGTGGCT		AAGAAAAA	AAAGATGTTCA	AAGAGAA
Calja3A21	GATCAAAA	AGAGATGTTCA	GGCCAGGTGCGGTGGCT		AAAAAAAA	ATAGATGTTCA	AAGAGAA
Calja3A5	GATCAAA-	-----------	-----------------		--------	AGAGATGTTCA	ACGAGAA
Calja3A90	GATCAAA-	-----------	-----------------		--------	AGAGATGTTCA	ACGAGAA
Homsa3A5	GATCAAA-	-----------	-----------------		--------	AGAGATGTTCA	AGGAGAA
Pantr3A5	GATCAAA-	-----------	-----------------		--------	AGAGATGTTCA	AGGAGAA
Macmu3A5	GATCAAA-	-----------	-----------------		--------	AGAGATGTTCA	AGGAGAA
Homsa3A43	GAT-AAA-	-----------	-----------------		--------	AGAGATAACGG	AGGGGAA
Pantr3A43	GAT-AAA-	-----------	-----------------		--------	AGAGATAACGG	AGGGGAA
Macmu3A43	AAT-AAA-	-----------	-----------------		--------	AGAGATAATGG	AGGGGAG
Otoga3A91	GATCAAG-	-----------	-----------------		--------	AGAGATAATTG	AGAGGAA
Otoga3A92							

Fig. 11. Diagnostic RGC insertions. (A) Phylogenetic distribution of the 54 analyzed RGCs. Filled circles mark the insertions of retroposons; triangles mark the presence of diagnostic indels. All insertions of retroposed elements and indels are assigned identifying number and letter codes. (B) Example alignment of the *CYP3A* sequences at the AluSx (retroposon 1) insertion site. Direct repeats are boxed. The length of the excised repeat region is indicated at the diagonal bars. Homsa (*Homo sapiens*), Pantr (*Pan troglodytes*), Macmu (*Macaca mulatta*), Papan (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

3.7 Absence of *CYP3A67* in humans

The designed primers successfully amplified from chimpanzee BAC DNA products of two distinct sizes representing *CYP3A67* and *CYP3A7*. All 137 tested human genomic DNA samples resulted in a single PCR band derived from *CYP3A7*. An example is shown in figure 12, where PCR products from 14 samples were separated on agarose gel.

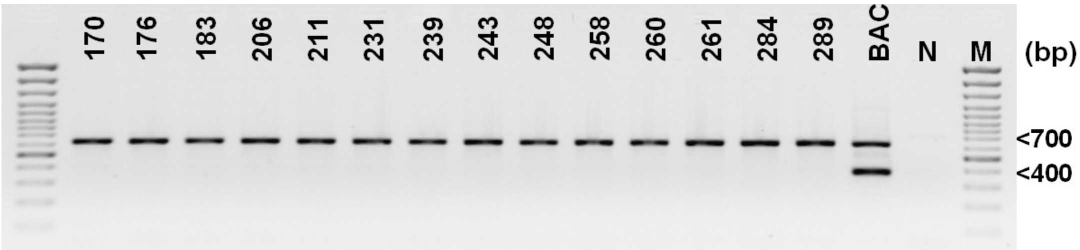


Fig. 12. Agarose electrophoresis of PCR for *CYP3A67* screening. Each lane was labeled with the ID number of the DNA sample from the NHL study. BAC: CH251-400N23; N: negative control; M: marker.

3.8 *CYP3A67* polymorphism in chimpanzees

PCR diagnostics of 5 BAC clones confirmed the two *CYP3A67* alleles in chimpanzee “Clint”, the primary DNA donor for CHORI-251 BAC library and the whole genome sequencing project (Chimpanzee Genome Sequencing Consortium. 2005). Three out of the five BAC clones appear to be derived from the deletion-containing *CYP3A67* allele (704-bp band), including the clone (CH251-35M15) which has been shotgun sequenced (Fig. 13). The only wildtype allele-containing clone is CH251-400N23 which resulted in a 1493-bp fragment. CH251-171D20 did not yield any PCR products, consistent with the fact that it only covers the distal part of the *CYP3A67* promoter as confirmed by shotgun sequencing. Of the four chimpanzee genomic DNA samples, one sample (“Lan”) resulted in a PCR product representing a deletion-containing *CYP3A67* allele but not the wildtype *CYP3A67*. The other three samples contained wildtype *CYP3A67* alleles (Fig. 14). Each PCR-amplified allele was verified by sequencing.

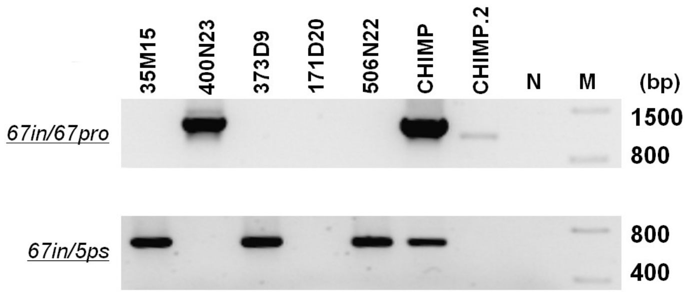


Fig. 13. PCR diagnostics of *CYP3A67* alleles in chimpanzee BAC clones. Each lane is labeled by the name of the BAC clone in library CHORI251. CHIMP and CHIMP.2 indicate different preparations of chimpanzee DNA samples. N: negative control, M: maker. The primer pairs used for genotyping are given on the left side of figure.

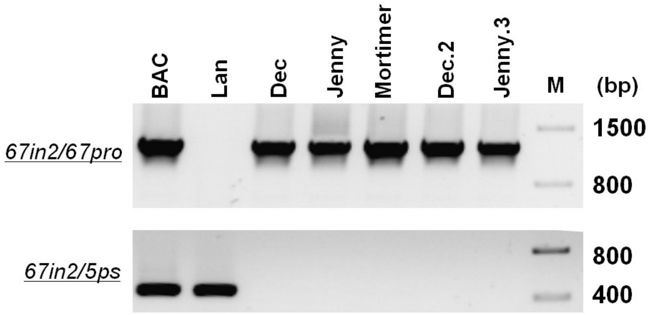


Fig. 14. PCR diagnostics of *CYP3A67* alleles in chimpanzee samples. Each lane is labeled by the name of the chimpanzee. Dec.2 and Jenny.3 indicate alternative preparations of the genomic DNA from the indicated individual. N: negative control, M: maker. The primer pairs used for genotyping are given on the left side of figure.

3.9 Evidence for positive selection among primate *CYP3A*

According to LRT I and II, the log likelihood values of discrete model M3 and beta&ω model M8 were significantly higher than the log likelihood value inferred under assumption of one-ratio model M0 and beta-null model M7, respectively (Table 6). Consequently, positive selection can be assumed for single sites of the 18 catarrhine *CYP3A* genes studied. According to parameter estimates from M3 and M8, about one-tenth of sites are under moderate positive selection ($\omega = 2.3$). Estimates from posterior probability provided significant support of positive selection for codons 437, 478 and 479 ($P \geq 0.95$), whether taking M3 (NEB) or M8 (NEB + BEB; Table 6) model. In case of codon 74, support from posterior probability was significant for M3 (NEB, $P \geq 0.95$) and nearly significant for M8 (0.94, NEB + BEB). Likewise, the free-ratio model fits the data significantly better than the null model, thus suggesting lineage-specificity of ω (LRT III in Table 6). Out of the 33 branches of the analyzed phylogeny, six branches showed evidence of positive selection (Fig. 15), with the highest ω values for branch Ho7 (CYP3A7 in the hominoid stem line; ω = infinite) and Hs4 (human *CYP3A4*; ω = infinite). The estimated numbers of nonsynonymous changes in Ho7 and Hs4 were 15.5 and 6.0, respectively, whereas the estimate of synonymous codon changes was 0 for either branch (Fig. 15). Pairwise comparison of the sequences at the basal and peripheral ends of Ho7 and Hs4 provided similar results. In detail, a total of 18 codon sites were found to contain nonsynonymous substitutions along Ho7 (H28R, S29T, V50A, F57Y, F74I, G77C, R78Q, V81M, L108F, S116N, S215P, I220V, S286T, V296M, E333K, A337T, N437G, R484L). The corresponding number for Hs4 was 6 (R54H, R78Q, I129L, I224T, R478S, T489V). The concurrent absence of synonymous exchanges renders these codon sites candidates of positive selection along Ho7 and Hs4. Notwithstanding these nonsynonymous exchanges, conservation has played a dominant role in the primate *CYP3A* evolution. Thus, slight to strong negative selection is detected for most sites (89%; Table 6). Moreover, most branches are characterized by ω values < 1 and, hence, negative selection.

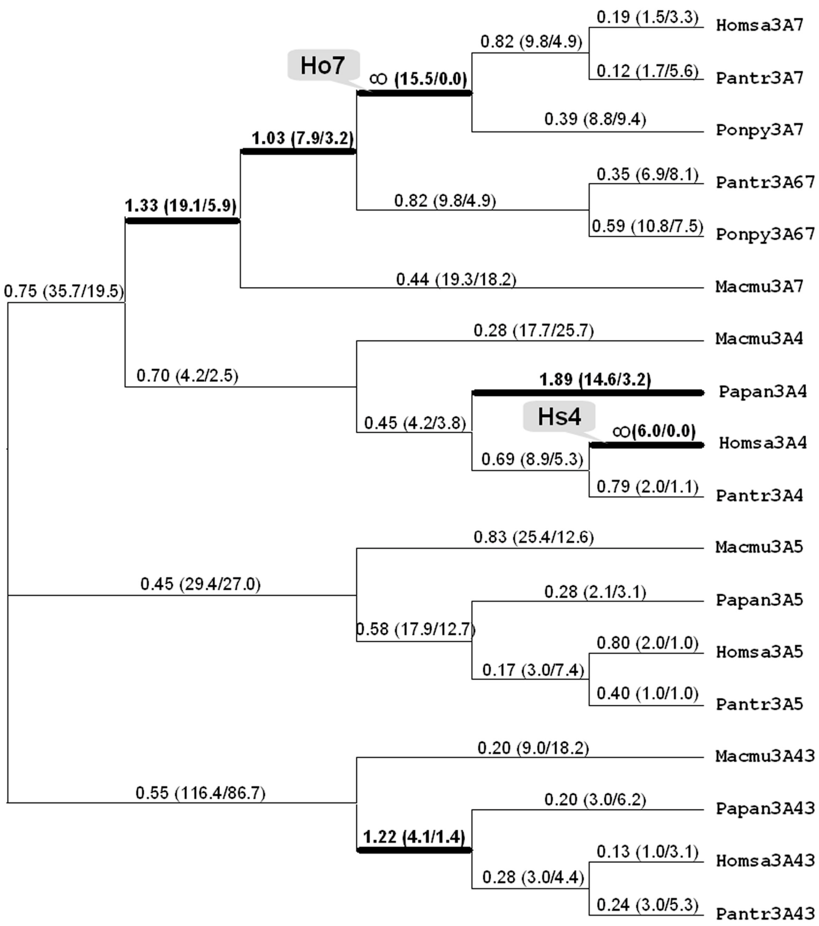


Fig. 15. Selection within catarrhine *CYP3A* phylogeny. The branches with ω ratios >1 as estimated by the free-ratio model are shown in thick lines. The estimated numbers of nonsynonymous and synonymous changes are given in the parentheses. Homsa (*Homo sapiens*), Pantr (*Pan troglodytes*), Macmu (*Macaca mulatta*), Ponpy (*Pongo pygmaeus*).

Table 6. Test statistics and parameter estimates for the selection in primate *CYP3A* genes

LRTs	Model	*L*	test statistics	np	Estimates of parameters	sites with highest support from BEB
LRT I (site-specific)	M1a (nearly neutral)	-5643.60	$2\Delta l = 11.44$ $cv = 9.21$	2	$p_0 = 0.56$, $(p_1 = 0.44)$ $\omega_0 = 0.10$, $(\omega_1 = 1.00)$	-
	M2a (positive selection)	-5637.88	$p < 0.01$	2	$p_0 = 0.60$, $p_1 = 0.33$, $(p_2 = 0.07)$ $\omega_0 = 0.15$, $(\omega_1 = 1.00)$, $\omega_2 = 2.63$	sites with $p_{\omega>1} = 0.87$-0.92: 437, 478, 479
LRT II (site-specific)	M7 (beta)	-5645.72	$2\Delta l = 15.78$ $cv = 13.82$ $p < 0.001$	2	$p = 0.15$, $q = 0{,}15$	-
	M8 (beta&ω)	-5637.83		4	$p = 0.50$, $q = 0{,}80$ $p_0 = 0.89$, $(p_1 = 0.11)$, $\omega = 2.29$	sites with $p_{\omega>1} \geq 0.95$: 437, 478, 479
LRT III (branch-specific)	M0 (one ratio)	-5692.18	$2\Delta l = 53.06$	1	$\omega = 0.54$	-
	free ratio	-5665.65	$cv = 46.19$ $p < 0.05$	33	see figure 15	-

BEB, Bayes empirical Bayes approach; cv, critical value from chi-square distribution given the difference in the number of free parameters of the models compared; np, number of free parameters; *L*, log likelihood; LRT I, likelihood ratio test for rate heterogeneity; LRT II, likelihood ratio test for the presence of positively selected sites; LRT III, likelihood ratio test for lineage-specificity of ω; $2\Delta l$, twice the log-likelihood difference of the models compared; p, level of significance.

3.10 Evolutionary analysis of primate *CYP3A* promoters

3.10.1 Characterization of primate *CYP3A* promoters

Consistent with the *CYP3A* phylogeny based on protein-coding sequences, high levels of similarity were limited to pairwise comparisons of orthologous promoters from different species or between *CYP3A4*, *CYP3A7/67*, and *CYP3A21* (Figs 16 and 17). Except for short stretches of homologous regions due to the aligning of repeat elements, there is almost no similarity when paralogs are compared (e.g., *CYP3A4* vs. *CYP3A43/5* or *CYP3A5* vs. *CYP3A43*). The promoters of the two marmoset *CYP3A5*-like (marmoset *CYP3A5* and *CYP3A90*) genes show very limited similarity to catarrhine *CYP3A5*. Compared with human and chimpanzee *CYP3A5*, there is a gap of approximately 5.5 kb in the rhesus *CYP3A5* promoter due to long interspersed elements (LINE) insertions in human and chimpanzee promoters. While galago *CYP3A* promoters display similarity with neither *CYP3A5* nor *CYP3A43* promoters, low level of similarity was observed when these two promoters were aligned with that of *CYP3A4/7/67*.

Despite of the lack of overall similarity among all primate *CYP3A* promoters, they are conserved in the proximal region that is up to ~850 upstream of the transcriptional start site. The conserved region extends to approximately -1300 bp when *CYP3A5* and *CYP3A5*-like promoters are excluded. Although *CYP3A4* and *CYP3A7/67* promoters were highly conserved at sequence level, large gaps emerged when *CYP3A7/67* were compared with *CYP3A4* promoters. In rhesus and baboon *CYP3A7* promoters, the high sequence homology extends up to ~12 kb, rather than to ~8.8 kb as do human and chimpanzee *CYP3A7* promoters. Moreover, a deletion of about 2.5 kb was found in *CYP3A67* in comparison to *CYP3A4*/*7* (Fig. 16).

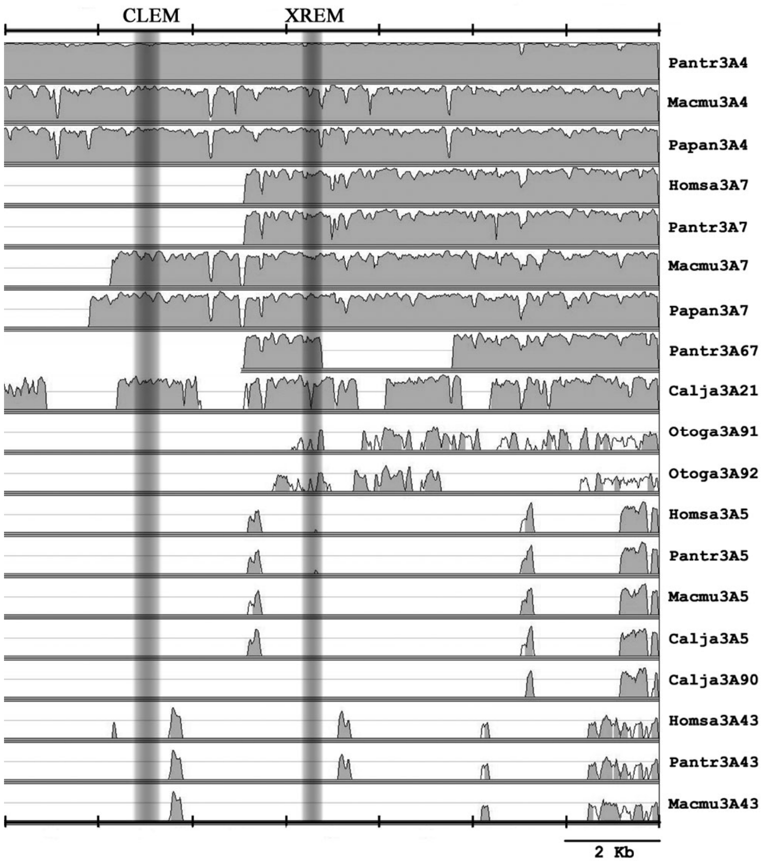


Fig. 16. Plot of identity between human *CYP3A4* and all the other primate *CYP3A* promoter sequences. The gaps in the reference sequence (*CYP3A4*) in all pairwise comparisons are not shown. The identity plot of *CYP3A67* promoter is incomplete because only partial sequence is available for it. Only regions with identity over 50% are depicted. The plot was displayed in Vista Browser (Mayor et al. 2000). Sequence identity was calculated using a sliding window of 100 bp. The locations of the CLEM and XREM are indicated.

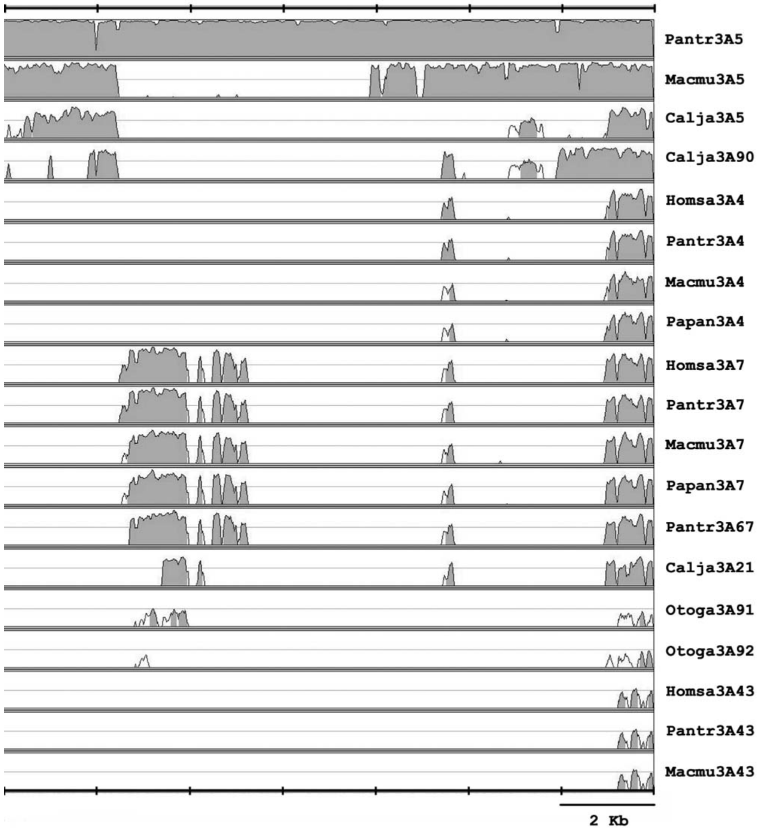


Fig. 17. Plot of similarity between human *CYP3A5* and all other primate *CYP3A* promoter sequences. The gaps in the reference sequence (*CYP3A5*) in all pairwise comparisons are not shown. All other details are as in Fig. 16.

3.10.2 Phylogeny of primate *CYP3A* promoters

Phylogenetic tree (Fig. 18A) was constructed for the 20 primate promoters based on a 618-bp sequence alignment of the conserved proximal promoter region (Fig. 16 and 17). The resulting tree is similar to the one obtained based on *CYP3A* coding regions (Fig. 10 and Fig. 18B), except that rhesus/baboon *CYP3A4* and *CYP3A7* form an OWM *CYP3A4/7* cluster and *CYP3A4/7/67* from human and chimpanzee are clustered in a separate hominoid clade. It is inconsistent with the gene tree in which *CYP3A4* and *CYP3A7* are separately clustered, irrespective of their species origins (from OWM or hominoid) (Fig. 10 and Fig. 18B). Although the promoter sequence-based tree is highly supported, the interior branches leading to OWM and hominoid *CYP3A4/7* clusters are short. Kishino-Hasegawa test shows that it does not fit the data significantly better than the coding sequence-based tree topology ($p < 0.16$). The phylogeny of *CYP3A4/7/67* based on the ~8.8 kb promoter sequences is consistent with the coding region based tree as shown in Figs. 10 and 18B.

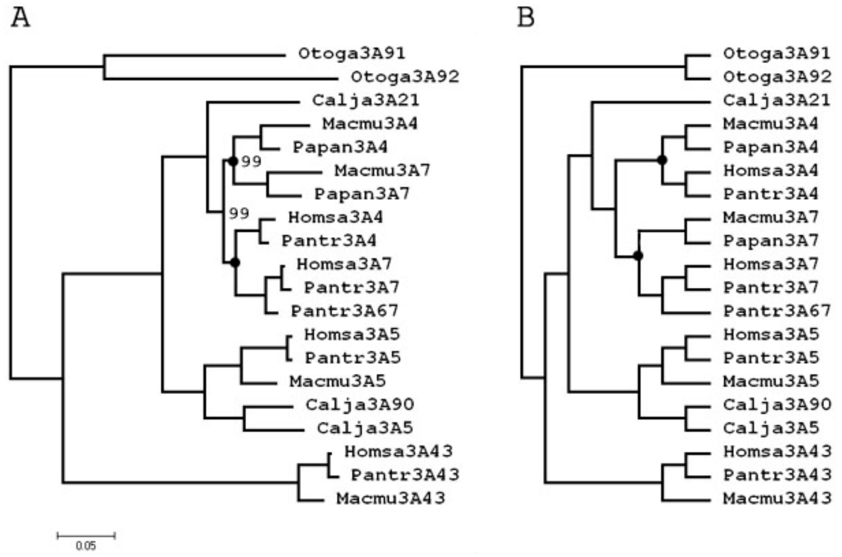


Fig. 18. The proximal promoter sequence-based tree topology (A) and coding sequence based tree (B). The nodes differing between the two trees are indicated by dots. The posterior probabilities lower than 100 are shown in (A). Homsa (*Homo sapiens*), Pantr (*Pan troglodytes*), Macmu (*Macaca mulatta*), Papan (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

3.10.3 Gene conversion among primate *CYP3A* promoters

GENECONV identified 5 possible recombination events between the *CYP3A4* and *CYP3A7* promoter regions from human, rhesus and baboon (Table 7 and Appendix Table 8.4). One of the involved fragments contains nuclear receptor response elements recognized by NHR-scan (Table 7). Additional evidence for gene conversion between rhesus *CYP3A4* and *CYP3A7* promoters is a 4 bp-deletion observed in only these two promoters (Fig. 19A). Another 4 bp-deletion found in chimpanzee, rhesus and baboon *CYP3A4*, and rhesus *CYP3A7* (Fig. 19B) suggests that the corresponding region has undergone independent gene conversion events in human and rhesus with opposite directions of sequence homogenization.

Table 7. Potential gene conversion events among *CYP3A4* and *CYP3A7* promoter regions.

Species	Position[1]	Length (bp)	Regulatory element[2]	No. of differences (bp)[3]	p value
Baboon	+58	289	ER6	6	0.00946
Rhesus	-867	145	/	0	0.03303
Human	-1501	274	/	5	0.01464
Rhesus	-1652	807	/	23	0.00001
Rhesus	-5373	111	/	0	0.00714

[1] Refers to the corresponding positions relative to human *CYP3A4* transcriptional start site. [2] Nuclear receptor binding sites detected by NHR-scan. [3] The difference between the two fragments involved in gene conversion. See Appendix Table 8.4 for details.

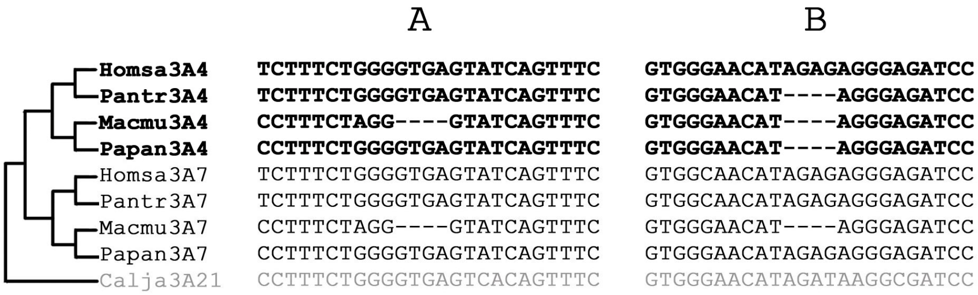

A
B
Homsa3A4 TCTTTCTGGGGTGAGTATCAGTTTC GTGGGAACATAGAGAGGGAGATCC
Pantr3A4 TCTTTCTGGGGTGAGTATCAGTTTC GTGGGAACAT----AGGGAGATCC
Macmu3A4 CCTTTCTAGG----GTATCAGTTTC GTGGGAACAT----AGGGAGATCC
Papan3A4 CCTTTCTGGGGTGAGTATCAGTTTC GTGGGAACAT----AGGGAGATCC
Homsa3A7 TCTTTCTGGGGTGAGTATCAGTTTC GTGGCAACATAGAGAGGGAGATCC
Pantr3A7 TCTTTCTGGGGTGAGTATCAGTTTC GTGGCAACATAGAGAGGGAGATCC
Macmu3A7 CCTTTCTAGG----GTATCAGTTTC GTGGGAACAT----AGGGAGATCC
Papan3A7 CCTTTCTGGGGTGAGTATCAGTTTC GTGGGAACATAGAGAGGGAGATCC
Calja3A21 CCTTTCTGGGGTGAGTCACAGTTTC GTGGGAACATAGATAAGGCGATCC

Fig. 19. Two indels in *CYP3A* promoter regions with non-canonical distribution in the primate *CYP3A* phylogenetic tree.

3.10.4 Analysis of potential nuclear receptor binding sites in primate *CYP3A* promoters

To explore the conservation of REs in primate *CYP3A* promoters, REs in 13 kb of human *CYP3A4*, *5* and *43* upstream promoter regions and their homologs in other primate *CYPP3A* were sampled. Note that only the human *CYP3A4*, *CYP3A5* and *CYP3A43* promoters analyzed are 13 kb in length, whereas the length of all the other prompters varies (Appendix Table 8.1), depending on the size of insertions and deletions when compared to their human homologs.

Altogether, 296 predicted REs belonging to 13 different categories have been identified in the 20 *CYP3A* promoters (Fig. 20). 26, 13 and 11 potential REs were detected in the 13 kb promoters of human *CYP3A4*, *CYP3A5* and *CYP3A43*, respectively. In *CYP3A4*, several well known, functionally important REs were assigned high score (> 4) by NHR-scanning, including the proximal ER6, the DR3 (dNR1) and ER6 (dNR2) of the XREM, and the ER6 of the CLEM (Fig. 20 and Fig. 21A). All these ERs are conserved in *CYP3A4* and *CYP3A7* from different Catarrhines. Notably, the proximal ER6, XERM ER6 and CLEM ER6 appear to be conserved even in the two galago genes (*CYP3A91*/*92*), which spilt early and have evolved totally differently. A large number of novel REs with low scores (< 2) were revealed throughout the *CYP3A* promoter sequences, as were some REs with intermediate scores (> 2 and < 4, Fig. 20 and Fig. 21A).

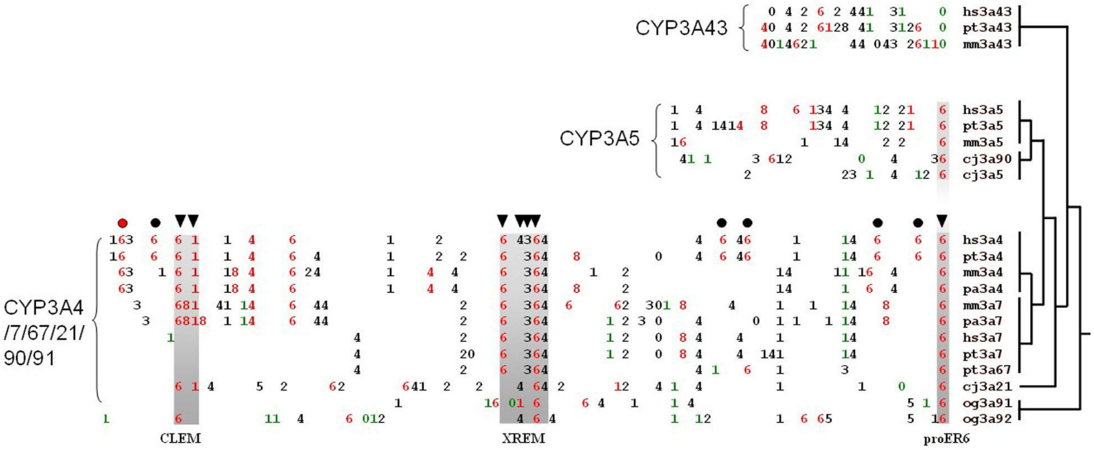


Fig. 20. Conservation of REs in *CYP3A* promoters. The digits indicate number of nucleotides which separate the two half-sites in each RE. Colors indicate the category of RE: red (everted repeat), green (inverted repeat) and black (direct repeat). The analysis comprised 13 kb of the upstream region of human *CYP3A4*/*5*/*43* and their homologous regions in other *CYP3A* promoters. The three known modules (proximal ER6, XREM and CLEM) are shadowed. REs in human *CYP3A4* with function confirmed experimentally are indicated by black arrows. Dots indicate the ER6 elements whose function is under investigation. ER6 sites which were likely generated in a common human/chimpanzee ancestor are indicated by black dots. hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

Nuclear receptor binding sites have been shown to be generated by repeat elements (e.g., primate-specific Alu) expansion (Laperriere et al. 2007). To explore the role of repeat elements in primate *CYP3A* regulatory element formation, locations of REs were compared with those of repetitive elements in primate *CYP3A* promoters. Altogether, 135 REs are contained in repeat elements while the other 161 REs are associated with non-repeat genomic sequence (Fig. 21B). Of the 26 REs identified in human *CYP3A4*, 14 REs are located in non-repeat element regions while the others are derived from ancient primate repeats (Appendix Table 8.7). Notably, the three best characterized *CYP3A4* regulatory modules (proximal ER6, XREM, and CLEM) all reside in non-repeat element regions. In contrast, nearly all REs in *CYP3A5* promoters are derived from repeat elements. Likewise, the 6 DR2 elements of the marmoset *CYP3A21* were all generated form repeat elements (Fig. 21B).

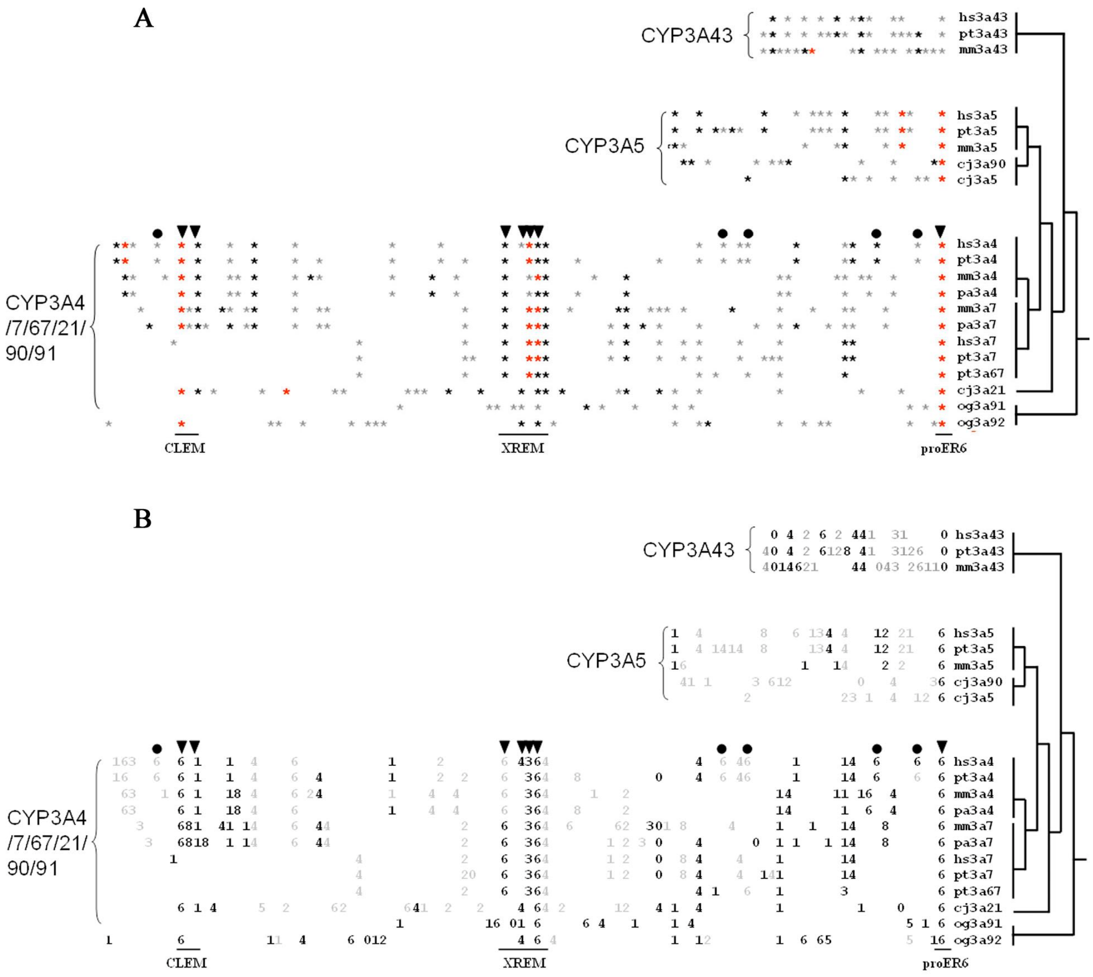


Fig. 21. REs scores in primate *CYP3A* promoters. (A) Grey: score <2, black: 2<=score<4, red: 4<=score; (B), REs within repeat elements are shown in grey. REs from non-repeats are black. REs in human *CYP3A4* with function confirmed experimentally are indicated by black arrows. ER6 sites which were likely generated in human/chimpanzee ancestor are indicated by black dots. hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

To explore the distribution of each category of REs in different promoters, the number of each of the three RE (ER, DR and IR, see Fig. 2 for details) was counted in the 13 kb promoter of each gene with the consideration of the spacer size. Contrary to the above RE conservation analysis, homology information was not taken into consideration. Because the length of sequence used for RE conservation analysis varies substantially (from ~7 kb to ~16 kb) and RE number tends to be higher in homologous sequences which contain large insertions and lower in sequences where deletions have occurred, the same size (13 kb) of sequence from each promoter was analyzed to enable unbiased comparison. *CYP3A67* promoter was excluded because only partial sequence was available. As the REs with intermediate (>= 2) and high scores (>= 4) fit well with the known functional regulatory elements, only REs with score not less than 2 were considered to reduce confounding effects due to false-positive prediction. Altogether, 129 potential REs belonging to 10 different RE categories were detected. In general, *CYP3A5* and *CYP3A43* contain fewer REs than *CYP3A4*/*7* (Fig. 22). The number of RE in each category ranged from 3 (DR0 and IR1) to 48 (ER6) (Fig. 23). The highest number was observed for ER6, followed by DR4 and DR2 (Fig. 23). ER6 elements are more numerous in *CYP3A4/7* than in *CYP3A5/43*. In *CYP3A4*/*7*/*21*/*91*/*92* promoters, ER6 number ranges from 2 in galago *CYP3A91/92* to 6 in human and chimpanzee *CYP3A4* (Fig. 24). DR4 is distributed almost evenly among all promoters (Fig. 24). An increased number of DR2 elements has been found in marmoset *CYP3A21* (Fig. 24).

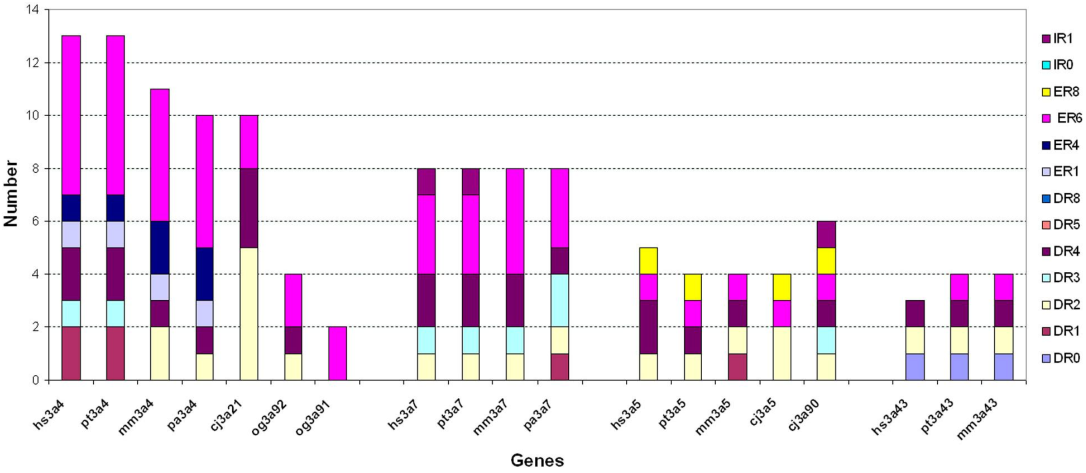


Fig. 22. Content of REs in primate *CYP3A* promoters. hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

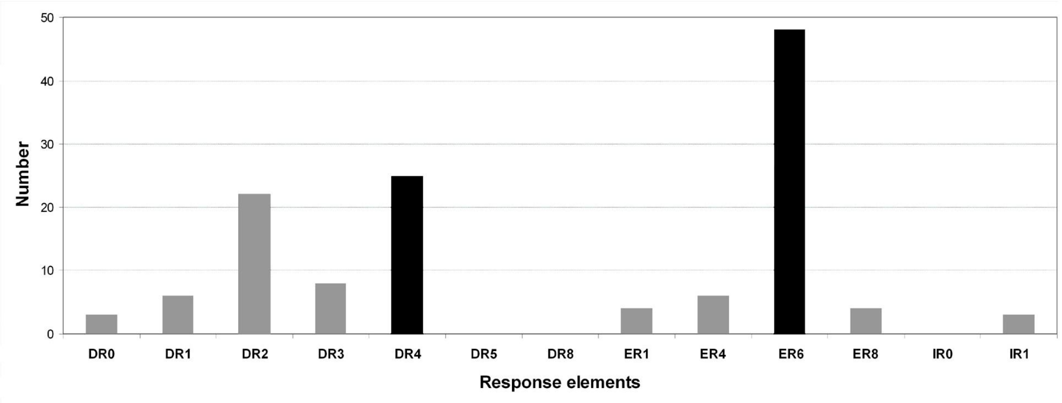


Fig. 23. The number of various REs in primate *CYP3A* promoters. The REs which display strong PXR-binding capacity (Frank et al. 2005) are depicted in black color. hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

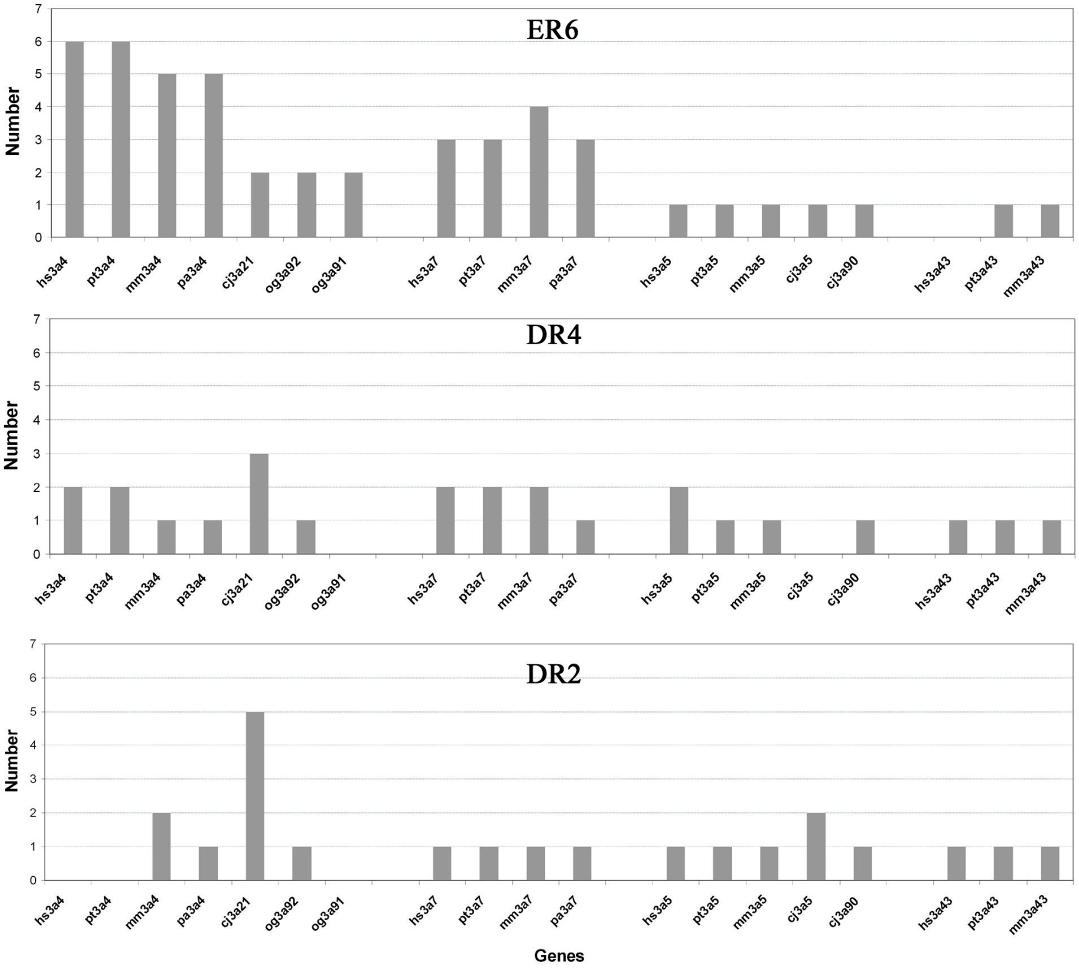


Fig. 24. The number of ER6, DR4 and DR2 elements in primate *CYP3A* promoters. hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).

4. Discussion

4.1 Comparative genomics and phylogenetics of vertebrate *CYP3*

The combination of comparative genomics, phylogenetics, and selection analysis described in this work appears to be a valuable extension of techniques usually applied to investigate enzyme function, such as X-ray analysis, spectroscopic measurements, site-directed mutagenesis, mechanism-based inhibition, and photoaffinity labeling. The last ~450 million years of *CYP3* evolution display several remarkable features. Most *CYP3* genes of the same clade in the phylogenetic tree built gene clusters within a genomic *CYP3* locus, suggesting a dominant role of tandem duplication in their development. Loci in distantly related species developed through independent gene duplications, despite striking similarities of their structures. In the individual loci, most *CYP3* genes are in a head-to-tail orientation, which confers higher stability of the locus (Graham 1995). The oppositely (head-to-head or tail-to-tail) oriented genes, if any, are always the most distantly related ones, both phylogenetically and physically. In some cases, these genes are located in another subclade of the phylogenetic tree (e.g. opossum *CYP3A80*) or even belong to a different *CYP3* subfamily (stickleback *CYP3D1*). This indicates a common role of inverted duplications in the early expansion of C*YP3* loci.

CYP3A genes from Amniota form two distinct groups in the phylogenetic tree ("*CYP3A37*" and "*CYP3A80*" in Fig. 7). We propose that at least two *CYP3A* genes existed in the Amniota ancestor before the split of Sauropsida and Mammalia. This is further supported by our comparison of the genomic structure of chicken, opossum, and platypus *CYP3A* loci. While opossum *CYP3A101-106* and platypus *CYP3A107-110* result from independent duplications of an ancestral *CYP3A37* homolog, the opossum *CYP3A80* gene arose from an ancestral *CYP3A80* homolog, which is in opposite orientation to the *CYP3A37*-related genes. As gene conversion played only a minor role in *CYP3* evolution, its confounding effects on the reconstructed phylogeny and functional divergence investigation was negligible.

Upon the emergence of eutherian mammals, the ancestral *CYP3A80* ortholog got lost, while that of *CYP3A37* translocated from *CYP3HR1* to *CYP3HR2*. This has led to a deserted *CYP3HR1* and to the evolution of *CYP3A* genes in eutherian mammals in an environment different from all other species. Interestingly, except of loss of *CYP3A* genes, eutherian *CYP3HR1* remains remarkably intact, indicating no structural instability. It is tempting to speculate that the novel genomic location, i.e. *CYP3HR2*, may have been selected following a rare translocation event from *CYP3HR1* due to certain fitness advantages. It is known that some genomic locations affect gene expression (Verschure 2004) and protein evolution (Pal et al. 2006; Williams and Hurst 2000). *CYP3A* is expressed in placenta, where it may play a fetus-protecting role by metabolising steroids and foreign substrates (Hakkola et al. 1998).

4.2 Clupeocephala *CYP3*: acquisition of novel subfamilies and functional divergence

The synteny of *CYP3* loci does not completely mirror the phylogenetic relationships among the analyzed representatives of Clupeocephala (Actinopterygii). While *CYP3HR1* typically harbors *CYP3A* genes, in zebrafish it contains four *CYP3C* genes with organ expression different from *CYP3A* (Corley-Smith et al. 2006) (Fig. 5). Conversely, the zebrafish *CYP3A65* is located apart from the *CYP3HR1*. A similar situation is observed in fugu and stickleback *CYP3* genes, where both *CYP3D* genes are found in *CYP3HR1*. Though CYP3D and CYP3C were classified separately from CYP3A due to reduced sequence similarity (< 55%) with CYP3A, these subfamilies actually represent extant CYP3A as judged from their locations in the most ancient *CYP3A* syntenic region found in vertebrates. In agreement, CYP3D is clustered together with CYP3A in the phylogeny tree (Fig. 25) although with long branches. One explanation is the accelerated mutation rate experienced by these two groups of CYP3 proteins (CYP3C and CYP3D), which is evident by the relative rate tests conducted.

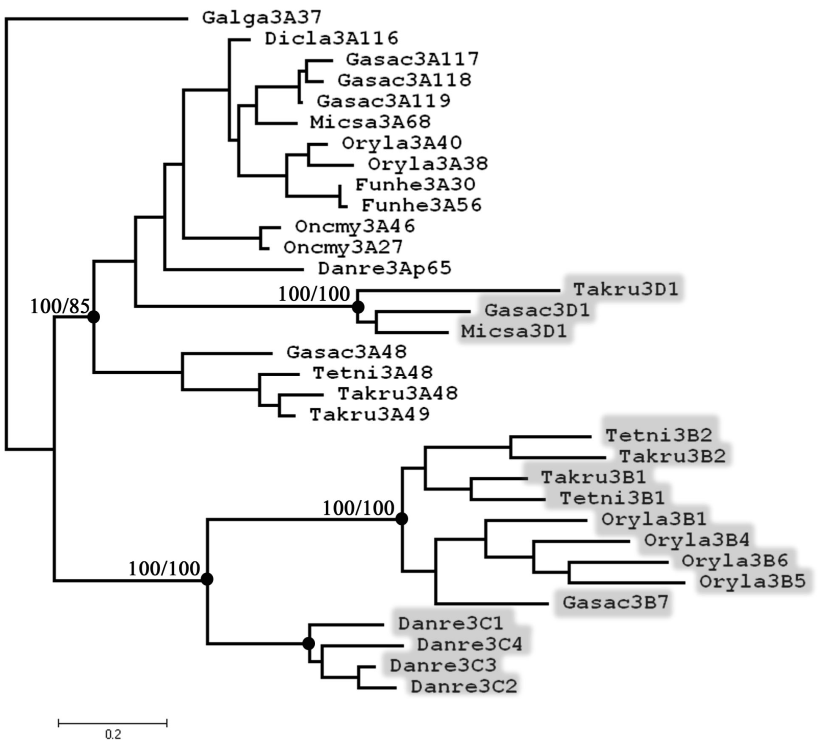


Fig. 25. Phylogeny of Clupeocephala CYP3 based on protein sequences. Critical nodes are labeled with black dots and their support values (posterior probability and bootstrap values). CYP3B/C/D genes are shadowed. Galga (*Gallus gallus*), Mga (*Meleagris gallopavo*), mgi (*Macropus giganteus*), Danre (*Danio rerio*), Talru (*Takifugu rubripes*), Oryla (*Oryzias latipes*), Gasac (*Gasterosteus aculeatus*), Tetni (*Tetraodon nigroviridis*), Dicla (*Dicentrarchus labrax*), Oncmy (*Oncorhynchus mykiss*), Funhe (*Fundulus heteroclitus*), Micsa (*Micropterus salmoides*).

Interestingly, the *CYP3C* genes which reside in *CYP3HR1* are unambiguously clustered with *CYP3B* genes located in another syntenic region, rather than with other *CYP3HR1*-dwelling *CYP3A* genes (Fig. 25). This discrepancy between the results from phylogenetic tree reconstruction and the conclusions drawn from comparative genomics can be explained by the following scenario (Fig. 26): *CYP3A*-containing *CYP3HR1* was duplicated during the whole genome duplication (WGD) which occurred in the stem lineage of teleost fish. *CYP3A* contained within one copy of the duplicated *CYP3HR1* underwent accelerated evolution resulting in *CYP3C* and in decreased similarity to other *CYP3A* proteins. *CYP3C* underwent further duplicative translocation event leading to the precursor of *CYP3B*, which subsequently experienced faster evolution as well.

Thus *CYP3B* and *CYP3C* were derived from one copy of *CYP3HR1* after WGD, whereas *CYP3A* and *CYP3D* originate from the other one. Examination of the available cDNA and EST data from several other Clupeocephala species (Appendix Table 8.5) indicates that

CYP3B/D and *CYP3C* must have existed at least in Percomorpha and Cypriniformes ancestors, respectively. It is further proposed that at least three subfamilies of *CYP3* (*CYP3A*, *B* and *C*) were already present in the Clupeocephala ancestor (the latest common ancestor of both zebrafish and Percomorpha), followed by a reciprocal loss of gene loci along the two resulting lineages, Euteleostei, leading to Percomorpha, and Otocephala which led, among others, to present day zebrafish. Although the exact time-point is unclear, *CYP3C* has been lost as early as in the Percomorpha ancestor, based on the absence of *CYP3C* genes in the genomes of four Percomorpha descendent lineages (takifugu, tetraodon, stickleback and medaka). The loss of *CYP3A* and *CYP3B* loci in Otocephala is hard to time, due to the lack of sufficient genome data from this lineage. Furthermore, accelerated evolution also happened to some copies of *CYP3A* in *CYP3HR1* which finally developed into extant *CYP3D* in Percomorpha. Altogether, the teleost fish-specific WGD appears to have provided starting material for the subsequent *CYP3* family expansion. Gene duplication and gene loss, accompanied by accelerated evolution and functional divergence (see sections 3.4 and 3.5), have played crucial roles in the evolution the *CYP3* family in Clupeocephala.

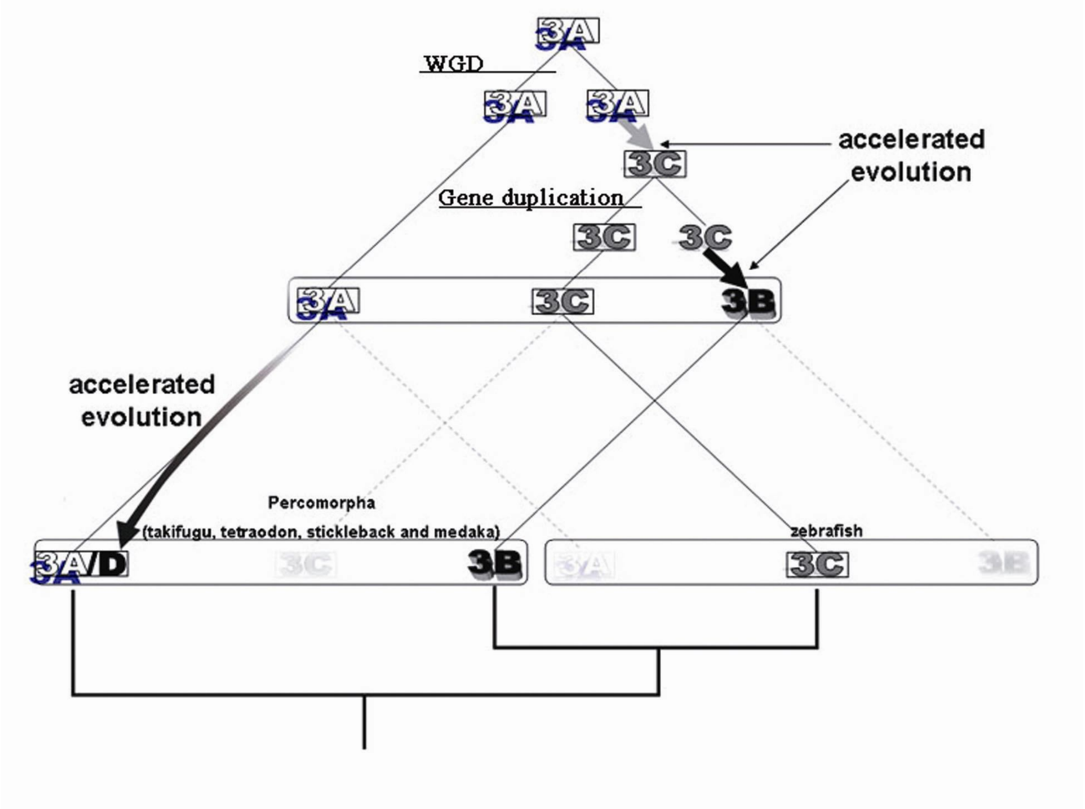


Fig. 26. *CYP3A-D* genes are represented by “3A”, “3B”, “3C” and “3D. The CYP3A/3C genes in the syntenic region *CYP3HR1* are framed in small boxes. The accelerated substitution events leading to *CYP3C*, *CYP3B* and *CYP3D* are indicated by the filled arrows. Gene loss events are indicated by grey dotted lines.

The result of present functional divergence investigation among CYP3 proteins is comparable to a previous analysis of a much smaller dataset (McArthur et al. 2003). The functional divergence is pronounced between land-living eutherian mammals and aquatic Clupeocephala as suggested by the high coefficients for all protein level comparisons between these two lineages (mammalian CYP3A vs. CYP3B/C and mammalian CYP3A vs. Clupeocephala CYP3A/D). It is consistent with the nonresponsiveness of the sea bass *CYP3A79* to typical mammalian *CYP3A* inducers (Vaccaro et al. 2007) which suggest that some Clupeocephala *CYP3A* genes are likely to be differently regulated compared to mammalian *CYP3A*. Moreover, our analysis reveals high functional coefficients (θ) for CYP3A/D and CYP3B/C indicating strong altered selective constraint after the CYP3A/D-CYP3B/C split. Thus, these two groups of genes likely fulfill different functions as exemplified by the expression of *CYP3C1* in the gill and skin of zebrafish rather than in liver or intestine (Corley-Smith et al. 2006). In addition, elevated substitution rate along lineages leading to CYP3B, C and D of Clupeocephala may result from adaptive selection, relaxation of functional constraints, or loss of function. However, all the three accelerated subfamilies are expressed according to EST and mRNA data (Barber et al. 2007; Corley-Smith et al. 2006), rather reflecting further functional divergence between CYP3A and CYP3D, and between CYP3C and CYP3B. Altogether, the high diversification of *CYP3* which has taken place in Clupeocephala species is suggestive of function diversifications in the adaptation to their specific aquatic environments.

It is usually assumed that amino acid substitutions leading to functionally divergent CYP3A proteins should mainly affect SRS regions (Gotoh 1992), which are important to their substrate specificity. In contrast, the sites that responsible for functional divergence between Clupeocephala CYP3 and mammalian CYP3A or between Clupeocephala CYP3A/D and CYP3B/C (see section 3.5, Fig. 8) were identified throughout the length of CYP3 sequences. Previous analysis only revealed functional divergence in the central part of CYP3 proteins encompassing SRS3 and SRS4 (McArthur et al. 2003). This might be due to the smaller dataset used compared with present study as the power of test increases with lager sequence number and sequence diversity. Similarly, non-SRS functional divergence sites or regions, were also identified by analysis of CYP2C family (da Fonseca et al. 2007). In humans, functional changes of CYP3A proteins due to polymorphism outside of any known functional important sites or regions have also been repeatedly reported (Human Cytochrome P450 Allele Nomenclature Committee, http://www.cypalleles.ki.se/). Moreover, there are many findings that non-SRS residues changes can also lead to functional changes of other P450 proteins (Domanski and Halpert 2001). For example, hydrophobic region N’ terminal region is thought to interact with membranes and may be important for entering of some substrates into the active site (Schleinkofer et al. 2005). Recently, non-SRS in F-G helix has also been shown to be important for CYP3A7 substrate specificities (Torimoto et al. 2006). Therefore, non-SRS residues, or regions, of CYP3 may fulfill much more functions that still need be investigated.

4.3 Evolution of primate *CYP3A* genes and loci

Primate *CYP3A* loci appear to have evolved very different in Strepsirrhini and New World monkey compared to Old World monkeys and Hominoidea. The head-to-head orientation of the galago *CYP3A* (*CYP3A91* and *CYP3A92*) resembles that of *CYP3A43* and *CYP3A4* in Catarrhini (Fig. 9 and Appendix Fig. 8.4). However, the sister group relationship of

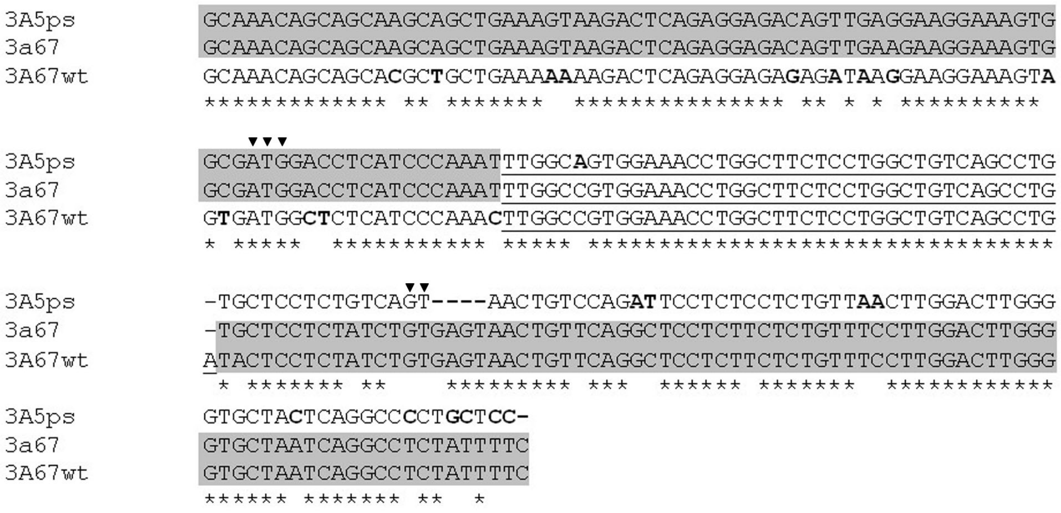
Strepsirrhini (i.e. galago) *CYP3A* and all anthropoid *CYP3A* in the reconstructed phylogeny suggests there was only one *CYP3A* gene in their common ancestor and all extant *CYP3A* in these two taxa arose through independent duplication along each lineage (Fig. 9 and 10). Contrary to the repeated duplications of *CYP3A21*-like genes (*CYP3A4*, *7*, *67*) in catarrhines, platyrrhines (New World monkeys) expanded only *CYP3A5*, as evidenced by the *CYP3A* locus structure of the marmoset. Evidence of a marmoset *CYP3A43* ortholog was found neither by screening a genomic BAC library, nor by the extension of the locus assembly based on the 6-fold coverage WGS data until the flanking gene *TRIM4* (data not shown). Since *CYP3A43* is located on the basal position of the anthropoid *CYP3A* phylogeny (Figs. 10 and 11), it must have been created prior to the split of catarrhines and platyrrhines, and therefore is likely lost in the latter species (marmoset). The small size and the relatively easy handling have led to the consideration of marmosets as a better model of human CYP3A pharmacology than rhesus (McArthur et al. 2003; Williams et al. 2004a). While *CYP3A21* and *CYP3A4* exhibit similarities in transcriptional regulation (Koehler et al. 2006), the 5’ upstream regions of marmoset notably differ from that of human *CYP3A5* due to independent retrotransposable element insertions. Taken together with the substantial differences between marmoset and catarrhine *CYP3A* gene sets, these data argue against marmoset being a good primate model of human *CYP3A*. The same is true for Strepsirrhini (i.e. galago) whose *CYP3A* genes evolved even more divergently.

4.4 Primate *CYP3A* gene loss

4.4.1 *CYP3A67* deletion in humans and polymorphic pseudogenization in chimpanzee

CYP3A67 arose through a duplication of *CYP3A7* early in the evolution of Hominidae, since it is found both in the chimpanzee and orangutan. Loss of *CYP3A67* is unlikely to be polymorphic in extant human populations, as judged from our negative, PCR-based screen of 99 Central Europeans and 38 African (Bantu) DNA samples. Due to the large gaps within the *CYP3A* locus in the first release of chimpanzee assembly (panTro1), the loss of *CYP3A67* in the human lineage had been thought to have occurred by homologous recombination between *CYP3A67* and the pseudogene downstream of *CYP3A7* (Rodriguez-Antona et al. 2005). In contrast, our data suggest an unequal crossover with the breakpoints located downstream of *CYP3A7* and *CYP3A67* (Appendix Fig. 8.5).

We also characterized an allelic deletion that encompasses *CYP3A67* promoter (Fig. 4). Considering the 4 chimpanzee sampled and the one subjected to whole genomic shotgun sequencing, 3 out of the 10 chromosomes appear to carry the deletion-containing allele. Thus the frequency of this polymorphic deletion may be non-negligible in chimpanzees. By comparing the flanking sequences of the deleted region with wildtype sequence, the breakpoint of the deletion was pinpointed to the first exon of the *CYP3A67* and *CYP3A5* pseudogene (Fig. 27). It is likely that a recombination event (probably unequal crossover) which occurred between these two exons subsequently led to the deletion of the intervening region and to a chimerical exon 1 (the *CYP3A67* exon 1 in the deletion-containing allele). These data suggest that the junction between *CYP3A67* and *CYP3A7* appears to be a hot spot of genomic rearrangements that led to both inter- and intra-species differences in *CYP3A* locus structure.



```
3A5ps           GCAAACAGCAGCAAGCAGCTGAAAGTAAGACTCAGAGGAGACAGTTGAGGAAGGAAAGTG
3a67            GCAAACAGCAGCAAGCAGCTGAAAGTAAGACTCAGAGGAGACAGTTGAAGAAGGAAAGTG
3A67wt          GCAAACAGCAGCACGCTGCTGAAAAAAAGACTCAGAGGAGAGAGATAAGGAAGGAAAGTA
                ************* ** *******  *************** ** * * **********

                   ▼▼▼
3A5ps           GCGATGGACCTCATCCCAAATTTGGCAGTGGAAACCTGGCTTCTCCTGGCTGTCAGCCTG
3a67            GCGATGGACCTCATCCCAAATTTGGCCGTGGAAACCTGGCTTCTCCTGGCTGTCAGCCTG
3A67wt          GTGATGGCTCTCATCCCAAACTTGGCCGTGGAAACCTGGCTTCTCCTGGCTGTCAGCCTG
                * *****  *********** ***** *********************************

                              ▼▼
3A5ps           -TGCTCCTCTGTCAGT----AACTGTCCAGATTCCTCTCCTCTGTTAACTTGGACTTGGG
3a67            -TGCTCCTCTATCTGTGAGTAACTGTTCAGGCTCCTCTTCTCTGTTTCCTTGGACTTGGG
3A67wt          ATACTCCTCTATCTGTGAGTAACTGTTCAGGCTCCTCTTCTCTGTTTCCTTGGACTTGGG
                 * ******* **    ********* ***  ****** *******  ************
3A5ps           GTGCTACTCAGGCCCCTGCTCC-
3a67            GTGCTAATCAGGCCTCTATTTC
3A67wt          GTGCTAATCAGGCCTCTATTTC
                ****** ******* **  *
```

Fig. 27. Sequence comparison of the first exon and its flanking regions in wildtype *CYP3A67* (3A67wt), *CYP3A5* pseudogene (3A5ps) and the deletion-containing *CYP3A67* (3a67) allele. Start codon of the exon 1 and splicing donor site (GT) are indicated by heads. The identical regions between sequences are grey-shadowed. Asterisks represent sites which are identical in all sequences compared. The region of the breakpoint is underlined.

Transcript of *CYP3A67* was considered to result from splicing of exon 1 of *CYP3A7* and exon 2-13 of *CYP3A67* (Williams et al, 2007). However, this conclusion appears to be invalid since it was based on the six mismatches between the cloned *CYP3A67* exon 1 and that of the *CYP3A67* from the genomic assembly (panTro1), which was mixed with the first exon *CYP3A5* pseudogene downstream of *CYP3A7*. The updated *CYP3A67* exon 1 (panTro2) differs from that of *CYP3A7* at only one site. Even this one nucleotide difference appears to be due to the polymorphism among individuals as judged from sequencing of our 4 chimpanzee samples. Thus, the first exons of *CYP3A67* and *CYP3A7* are identical in some or most chimpanzees. To test the utilization of *CYP3A7* exon 1 by *CYP3A67*, genomic and cDNA sequences of *CYP3A67* exon 1 should be compared from chimpanzees that carry polymorphisms which distinguish *CYP3A7* exon 1 sequence form that of *CYP3A67*. Comparison of orangutan *CYP3A67* exon 1 derived from cDNA sequencing and genomic DNA suggested no analogous trans-splicing event between *CYP3A7* and *CYP3A67* in this species (data not shown).

4.4.2 *CYP3A43* pseudogenization

Similarly to human *CYP3A43* (Domanski et al. 2001a; Gellner et al. 2001), chimpanzee *CYP3A43* produces mostly aberrant transcripts (data not shown). In addition, *CYP3A43* is likely to have become a pseudogene in some rhesus populations. The *CYP3A43* sequences in the WGS database of this species contain a point deletion in exon 9 which results in a premature stop codon 222 bp downstream of the deletion (Trace Identifier: 545318077, 497057015, 352000725 and 486641435). We found the same deletion in one of the three rhesus samples investigated (Fig. 28). These observations are consistent with an ongoing pseudogenization of *CYP3A43*, although a protein-independent function (Hirotsune et al. 2003) cannot be excluded. In support of the latter possibility, the coding region of *CYP3A43* appears to have evolved under purifying constraint, as evidenced by the low ω values for the terminal branches leading to these genes (Fig. 15).

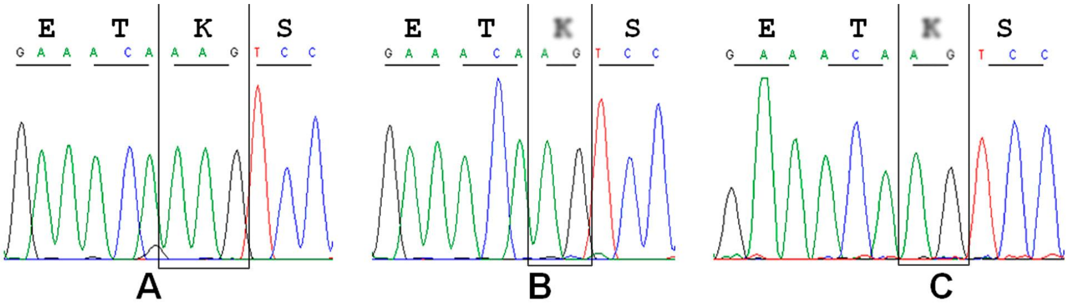


Fig. 28. DNA sequencing chromatogram of rhesus *CYP3A43* polymorphic indel. (A) read of wildtype *CYP3A43* cDNA sequence, (B) read of *CYP3A43* mutant cDNA sequence, (C) read of *CYP3A43* mutant genomic DNA sequence from Trace Archive (Trace Identifier: 545318077). Codons are underlined and the encoding amino acids are shown in a single letter code. The codon affected by the single base polymorphic deletion is boxed.

Due to the multitude of possible mechanisms, gene loss or pseudogenization are more likely outcomes of a mutation than of gain-of-function gene variants (Olson 1999). The loss of *CYP3A67* (humans) and *CYP3A43* (marmoset), and the pseudogenization of *CYP3A43* (human, rhesus and chimpanzee), may therefore represent adaptive responses. Taken together, the evolution of primate *CYP3A* was a complex process involving both gene birth and death in a lineage-specific manner.

4.5 Positive selection and its functional implications

The presented analysis of a set of primate *CYP3A* genes revealed a dominant role of purifying (i.e. negative) selection in the primate CYP3A protein evolution, consistent with the conservation of their original biochemical functions. In addition, two episodes of accelerated sequence evolution were detected along branches Ho7 and Hs4 (Fig. 15). In general, variation of ω among lineages is not considered to be a sufficient evidence for adaptive evolution, since it can as well result from relaxation of selective constraint (Yang and Bielawski 2000). However, in case of Ho7 and Hs4 evidence for positive selection is rather strong: First, the ω estimate for Ho7 and Hs4 is infinite, i.e. as high as it can be. Second, the estimated numbers of nonsynonymous (Ho7: 15.5, Hs4: 6) and synonymous exchanges (Ho7: 0, Hs4: 0) are too different to explain the ω estimates by stochastic variation. Third, the results from branch-specific analysis (free ratio model) are supported by pairwise sequence comparisons which identified 18 and 6 candidate sites of positive selection along Ho7 and Hs4, respectively. Taken together, it appears justified to postulate positive selection for Ho7 and Hs4, instead of relaxation of the selective constraint. The selection of *CYP3A7* took place simultaneously with the origin of Hominoidea, or with the split of this superfamily into Hominidae (human, chimpanzee, gorilla, and orangutan) and Hylobatidae (gibbon). The exact evolutionary branch will be identified only after the cloning of gibbon *CYP3A* genes. Around the same time period, a gene conversion replaced exon 6 of *CYP3A7* with that of *CYP3A4* and *CYP3A7* became a predominantly fetal gene in hominoids. This latter conclusion is based on the absence of *CYP3A7* expression in adult hominids, in contrast to rhesus, olive baboon, and hamadryas baboon. All these events are consistent with a rapid change of an existing one, or with the acquisition of an entirely new function by CYP3A7, as indicated by the suppression of its expression in human adult livers, which might have been caused by the loss of its CLEM and the disruption of proximal ER6.

Strong positive selection acted also on *CYP3A4* following the split from the chimpanzee lineage (branch Hs4, Fig. 15).

Of the 18 amino acids identified as positively selected in CYP3A7, all but residues 78 and 108 differ between the contemporary human CYP3A4 and CYP3A7 protein sequences. 78Q is also the only amino acid common for CYP3A4 and CYP3A7 out of 6 residues positively selected in the human CYP3A4. Apparently, 78Q arose from an ancestral R twice, during the independent selection episodes on branches Ho7 and Hs4. This parallel amino acid exchange (R78Q) suggests a particular functional advantage of 78Q, which is at present unknown.

108F is found in both Hs CYP3A4 and CYP3A7, yet a strong selection signal is detected only in branch H7. This apparent contradiction can be explained by the plesiomorph F residue, common for the CYP3A21/CYP3A4/CYP3A7/CYP3A67 group, first changing into 108L within the CYP3A7 lineage and reverting to 108F due to the selection on Ho7. Together with 215F and 5 other phenylalanine residues, 108F forms a highly ordered hydrophobic core above the active site of CYP3A4, which may be involved in the initial recognition of substrates or allosteric effectors, or interact with the electron donors cytochrome b5 or P450 reductase (Williams et al. 2004b). The identity of the amino acid at position 108 in CYP3A4 affects the activity towards midazolam (Khan et al. 2002), aflatoxin B1 (Wang et al. 1998), the steroids testosterone and progesterone (Domanski and Halpert 2001; Wang et al. 1998), and lapachole (Wen et al. 2005). The importance of the 215P, selected in branch Ho7, for hominoid CYP3A7 proteins, is more difficult to ascertain. Together with amino acids 108 and 120, phenylalanine at position 215 plays a role in the binding and orientation of progesterone in the CYP3A4 active site (Park et al. 2005), but its mutagenesis affected neither the activity nor the cooperativity of the enzyme (Domanski et al. 2001b; Harlow and Halpert 1997).

Of the 18 amino acids selected in hominid CYP3A7 (Ho7), 8 are located within the first 100 out of 503 amino acids of the protein. This unusual clustering may affect the enzyme's localization to target structures such as endoplasmic reticulum or mitochondria, which depends on the N terminus. In addition, the N terminus affects the interaction with redox partners such as P450 reductase and cytochrome b5 (Domanski and Halpert 2001). In contrast to CYP3A4 and CYP3A5, CYP3A7 has been suggested to be less dependent on b5, although this may be a substrate-specific rather than general feature (Yamaori et al. 2003). The exact molecular determinants of the reduced interaction with b5 are unknown. The hydrophobic region of the N terminus may be important also for entering of some substrates into the active site (Schleinkofer et al. 2005).

Codons 74, 437, 478, and 479 have been identified as undergoing positive selection across the whole phylogeny. The replacement of the CYP3A4 478S with 478D (like in CYP3A5) had no effect on testosterone hydroxylation, but it reduced by 80-90% the activity and changed the regioselectivity towards aflatoxin B1 metabolites (Wang et al. 1998). The substitution of CYP3A4 L479 with T479 (like in CYP3A5) had a similar, reducing effect on the production of the same AFB1 metabolites. Additionally, it halved the hydroxylase activity of the mutant protein towards testosterone. The mutation of the CYP3A4 479L to 479F (like in CYP3A7) changed the product profile of 7-hexoxycoumarin (Khan and Halpert 2000). The introduction of CYP3A4-derived L at position 479 of the rat CYP3A9 decreased imipramine N-demethylation and steroid hydroxylation. The latter effect was modified by the identity of the neighboring residue 480 (Xue et al. 2003). Taken together, three (108, 478, 479) of the four amino acid residues identified here as positively selected affected the catalytic activity and/or regioselectivity of the enzyme in a complex, substrate-specific manner, when investigated by mutagenesis. The data on residue 215 is less clear.

At present, we can only speculate about the physiological importance of the two branch-specific positive selection events. Adaptation to global environmental changes is less likely than changes in homeostasis or diet, since both selection events were lineage- rather than geographic region-specific. Human CYP3A7 is the only CYP3A expressed in fetal liver and it displays pronounced differences in the metabolism of endogenous substrates, e.g., testosterone 6β-hydroxylation (Ohmori et al. 1998) and DHEAS 16α-hydroxylation (Kitada et al. 1987), compared with CYP3A4, the dominant isoform expressed in the adult liver. The amount of DHEAS in the brain exceeds that in the adrenals, spleen, kidneys, testes, liver and plasma (Leowattana 2004). Livers of human anencephalic fetuses exhibit extremely low expression of CYP3A7 (Leeder et al. 2005), altogether suggesting a role of CYP3A7-mediated DHEAS metabolism in brain development. Compared to apes, which are mostly herbivorous, humans consumed increasing amounts of animal foods during the last 2 million years (Leonard 2002; Milton 2003). The use of fire (i.e. cooking) may have further changed the composition of human diet, which may have accelerated the evolution of human CYP3A4. The resulting changes of the protein sequence may have brought about the wide CYP3A4 substrate spectrum observed in humans. The investigation of different catalytic functions between human and chimpanzee CYP3A4 is ongoing. Interestingly, positive selection has been recently detected in the ligand-binding domains of the *CYP3A* regulators PXR and CAR, suggesting adaptation to changing exposure to environmental toxins (Krasowski et al. 2005).

4.6 Characterization of primate *CYP3A* promoters

As in *CYP3A* genomic sequences, gene conversion (5 fragments) was also detected in *CYP3A* promoter regions. It may partially explain the discrepancy between the tree topologies that derived from the proximal promoter sequences and *CYP3A* genomic sequences (Fig. 18), since the *CYP3A4/7* proximal promoter regions from both rhesus and baboon contain fragments that are involved in gene conversion (Table 7). The two additional potential gene conversion events (Fig 19), not detected by GENECONV, but evidenced by indels with distribution inconsistent with the *CYP3A* gene phylogeny, suggest that there may have been more undetected conversion events. However, as there is no consensus method to date for gene conversion detection and different methods can produce very different results (Posada and Crandall 2001; Posada and Crandall 2002), the precise detection of gene conversions in *CYP3A* promoters and exploration of their probable roles in *CYP3A* expression is difficult.

Among the four human *CYP3A* promoters, *CYP3A4/7* promoters show highest similarity to the galago *CYP3A* promoter, suggesting a conservation of the original regulatory elements. In contrast, the promoters of *CYP3A5/43* were intensively shaped by genomic rearrangements (i.e., repeat element insertion and recombination). Consistent with the conservation of the proximal promoter region across all *CYP3A* at sequence level as shown (Fig. 16 and 17), proximal ER6 was detected in all *CYP3A* studied with the exception of *CYP3A43* (Fig. 20). It is due to several mutations accumulated in the proximal ER6 motif of *CYP3A43* (Fig. 29). Due to the prominent function of ER6 in *CYP3A* expression and induction, the disruption of *CYP3A43* ER6 is in agreement with its pseudogenization indicated by a low level of mostly aberrant transcripts.

GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Homsa3A4
GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Pantr3A4
GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Macmu3A4
GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Papan3A4
GAATA**T**T**AACT**CAA**T**GG**AGGTCA**GTGAG	Homsa3A7
GAATA**T**T**AACT**CAA**T**GG**AGGTCA**GTGAG	Pantr3A7
GAACA**TGAACT**CAAAG**CAGGTCA**ATGAG	Macmu3A7
GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Papan3A7
GAATA**TGAACT**CAAAGG**AGGTCA**GTGAG	Pantr3A67
GA--A**TGAACT**CAAAGG**AGGTCA**GTGGG	Calja3A21
GAATA**TGAACT**CAAA**AAAGGTCA**GTATA	Otoga3A91
TAATA**TGAACTTT**AAGG**AGGTCA**ATGTG	Otoga3A92
GAACA**TGAACT**CAAA**A**G**AGGTCA**GCAAA	Homsa3A5
GAACA**TGAACT**CAAA**A**G**AGGTCA**GCAAA	Pantr3A5
GAACA**TGAACT**CAAAGG**AGGTCA**GCAAA	Macmu3A5
GAATA**TGAACT**CA**C**AGG**AGGTCA**GCAAA	Calja3A5
GAATA**TGAACT**CA**C**AGG**AGGTCA**GCAAA	Calja3A90
GAATAG**GAA**A**T**CAAAGG**AGG**C**CA**AAATA	Homsa3A43
GAATAG**GAA**A**T**CAAAGG**AGG**C**CA**AAATA	Pantr3A43
GAATAG**GAA**A**T**CAAAGG**AGG**C**CA**AAATA	Macmu3A43

Fig. 29. The sequence alignment of the conserved proximal ER6 element in primate *CYP3A* promoters. The two 6-nucleotice half sites and substitutions in the spacer are shown in bold. Substitutions within the half sites are shadowed.

The NHR-scan prediction of REs in human *CYP3A4* identified most known functional sites in the three major regulatory modules (i.e. proximal ER6, XREM and CLEM). Although HNF4α is known to predominantly bind to DR1 and DR2 elements (Fraser et al. 1998; Jiang et al. 1995), two known functional HNF4α binding sites were found to overlap with an ER6 element in XREM and an ER1 in CLEM, respectively (Fig. 3). In addition, a predicted XREM DR4 element is contained in a rifampin-responsive region as confirmed in DNase I foot-printing assay (Goodwin et al. 1999). Taken together, of the 12 nuclear receptor binding sites/regions with evidence for function, 7 were successfully identified by NHR-scanning (Fig. 3). Notably, almost all high score REs are restricted to the three functional modules and overlap with the functional sites, suggesting the high specificity of the prediction.

A number of REs conserved in all four hominoid *CYP3A7* promoters were detected in only one or two of the four *CYP3A4* promoters (Fig. 20). This might reflect the different constraint under which these REs evolved in *CYP3A7* and *CYP3A4* promoters. Absence of CLEM was found in human and chimpanzee *CYP3A7*. As CLEM is conserved in Old World Monkey (rhesus and baboon), its loss must have occurred recently, most likely in a hominoid ancestor. Since CLEM is critical for constitutive expression of *CYP3A4*, the loss of *CYP3A7* CLEM, in addition to the disruption of the proximal ER6, may be causative to the absence of hepatic *CYP3A7* expression in most human adults. Consistent with it, *CYP3A7*, with its both proximal ER6 and CLEM well preserved, is expressed in the livers of adult Old World Monkeys. The genomic deletion of CLEM and of its surrounding region in human and chimpanzee is likely to be mediated by a LINE insertion, as the corresponding region is replaced by large stretch of L1PA5 3'-half sequence. Although chimpanzee *CYP3A67* promoter contains a large deletion (Fig. 16), its XERM is preserved. As only partial sequence of the *CYP3A67* promoter sequence is available, it is unknown whether or not CLEM is present in the *CYP3A67* distal promoter.

Except for the proximal ER6, no conserved REs were found in the entire *CYP3A5*-like gene promoter set, due to the very limited sequence similarity among them. Different *CYP3A5* promoters in fact consist of very different genomic contents. The extant catarrhine *CYP3A5* promoters are largely composed of the disrupted anciently *CYP3A5P1* wreck (Fig. 1 and Appendix Fig. 8.2), whereas the promoters of two marmoset *CYP3A5*-like genes were mainly derived from repeat elements insertions (Appendix Fig. 8.3).

Consistent with the high content of repeat elements in human *CYP3A*, REs derived from repeat elements account for nearly half of the REs in the promoter regions investigated. Six of the 11 human *CYP3A4* ER6 and 3 of the 5 human/chimpanzee-specific ER6 elements overlap with repeat elements (Fig. 21B). Interestingly, an AluJo repeat (-3.5 kb) alone contains three human/chimpanzee specific REs (two ER6 and one DR4) (Appendix Table 8.7). In agreement with the role of Alu in DR2 expansion (Laperriere et al. 2007), 5 out of the 6 *CYP3A21* DR2 overlapping with repeat elements are derived from Alu repeats (data not shown). Compared with REs derived from repeat elements, the three known *CYP3A4* modules possess more ancient origins, with their homologous regions being detectable in other non-primate mammalian *CYP3A* promoters (data not shown). Thus, repeat elements appear to be a substantial recent source of REs in *CYP3A* promoters, which might be responsible for gene- and species- specific regulation of *CYP3A*.

Considering the density of different REs (with intermediate or high scores) in primate *CYP3A*, the enrichment of ER6 and DR4 is consistent with the roles of CYP3A in xenobiotic metabolism, as ER6 and DR4 elements are frequently targeted and bind with high affinity by xenosensors such as PXR and CAR (Frank et al. 2003; Frank et al. 2005).

However, only ER6 number varies substantially (from 1 to 6) in different *CYP3A* promoters (Fig. 24). Human and chimpanzee *CYP3A4* display the highest density of ER6 in the proximal 13 kb promoter region. The ER6 density in other *CYP3A* promoters decreases with the increase of their phylogenetic distance to human/chimpanzee *CYP3A4*. Thus there seems to be a trend towards increase in ER6 elements along the lineage leading to human/chimpanzee *CYP3A4*. In comparison to *CYP3A4*, the lower density of ER6 in *CYP3A7*/*21*/*90*/*91* promoters is partially due to deletions of the homologous ER6 containing genomic regions or repeat element insertions that diluted ER6 density in the 13-kb promoters. Human and chimpanzee *CYP3A4* acquired a new ER6 in a region (-0.9 kb) which is conserved across all *CYP3A4*/*7*/*21*/*90*/*91*. Besides, the high number of DR2 elements observed in primate *CYP3A* promoters may contribute to the high expression of *CYP3A* in liver, because DR2 elements are binding sites for liver-enriched nuclear receptor, HNF4α (Fraser et al. 1998; Jiang et al. 1995). The impact of the particularly high number of DR2 elements in marmoset *CYP3A21* (Fig. 24) promoter requires further investigation.

The increase of ER6 number in human and chimpanzee *CYP3A4* promoter is even more pronounced when all REs in *CYP3A* promoter are considered without any constraint on prediction scores. Altogether, five ER6 elements can be identified, which are likely to have been generated in a human/chimpanzee ancestor in the *CYP3A4* promoter (Fig. 20 and Appendix Fig. 8.6), whereas only one ER6 element was found to be specific to rhesus and baboon *CYP3A4* (Fig. 20 and Appendix Fig. 8.6). The sequences of the ER6 containing regions are well conserved across homologs (Appendix Fig. 8.6). All these ER6 elements arose by one or two nucleotide substitutions within ER6 precursors, rather than being introduced by genomic insertions in human chimpanzee ancestor. The high number of ER6 and DR4 in *CYP3A* promoters might be responsible for the overall high inducibility of *CYP3A* by xenobiotics, whereas the rapid ER6 increase in human and chimpanzee *CYP3A4*, which might have been caused by positive selection of ER6 over other ER types, further led to even higher inducibility of human *CYP3A4* and its variability among individuals. Test of the functionality of the 5 human/chimpanzee-specific ER6 elements is ongoing.

The above observations, together with the positive selection that acted on the CYP3A4 protein, allow hypothesizing about the reason for the large inter-individual differences in the expression of *CYP3A4* observed among contemporary humans. Accelerated protein evolution is accompanied by changes (frequently increases) in the gene’s expression levels

(Nuzhdin et al. 2004) (Lemos et al. 2005), since both processes can improve fitness. For example, an increased activity of an enzyme such as CYP3A4 could be brought about through structural changes to the protein or by its increased expression. There is increasing evidence for a coupling between the evolution of protein sequence and their expression levels. On the other hand, the neutral theory of evolution predicts that the variation in gene expression among individuals of a species increases in parallel to the divergence between species (Khaitovich et al. 2006). Taken together, the interindividual variability in *CYP3A4* expression in humans may reflect the accelerated evolution of human *CYP3A4* regulatory elements following the split of human and chimpanzee lineages. In agreement with this hypothesis, *CYP3A4*, which displays the largest interindividual expression variability of all human CYP3A proteins, is the only *CYP3A* that shows a recent positive selection of its protein sequence.

5. Abstract

CYP3A metabolize 50% of currently prescribed drugs and are frequently involved in clinically relevant drug interactions. The understanding of roles and regulations of the individual *CYP3A* genes in pharmacology and physiology is incomplete. Using genomic sequences from 16 species we investigated the evolution of *CYP3* genomic loci over a period of 450 million years. Novel *CYP3* subfamilies (*CYP3B*, *C* and *D*) developed from their *CYP3A* precursors in Clupeocephala species through accelerated evolution. Pronounced functional divergence occurred between mammalian *CYP3A* and Clupeocephala *CYP3*. All amniota *CYP3A* genes evolved from two ancestral *CYP3A* genes. Upon the emergence of eutherian mammals, one of them was lost while the other acquired a novel genomic environment due to translocation. In primates, *CYP3A* underwent rapid evolutionary changes involving multiple gene duplications, deletions, pseudogenizations, and gene conversions. The expansion of *CYP3A* in catarrhines (Old World monkeys, great apes and humans) differed substantially from New World primates (e.g. common marmoset) and strepsirrhines (e.g. galago). We detected two recent episodes of particularly strong positive selection acting on primate *CYP3A* protein-coding sequence: (1) on *CYP3A7* early in hominoid evolution, which was accompanied by a restriction of its hepatic expression to fetal period, and (2) on human *CYP3A4* following the split of the chimpanzee and human lineages. In agreement with these findings, three out of four positively selected amino acids investigated in previous biochemical studies of CYP3A affect the activity and regioselectivity suggesting CYP3A7 and CYP3A4 may have acquired catalytic functions especially important for the evolution of hominoids and humans, respectively. Characterization of primate *CYP3A* promoters revealed an enrichment of ER6 elements in primate *CYP3A* promoters and a trend towards increase in ER6 formation along lineages leading to human and chimpanzee *CYP3A4*. The increase in the number of ER6 elements may be causative to the pronounced *CYP3A4* inducibility and expression variability in humans.

Zusammenfassung

CYP3A-Enzyme verstoffwechseln mehr als 50% aller gegenwärtig in der Therapie eingesetzten Wirkstoffe und sind daher häufig an klinisch-relevanten Arzneimitttel-Wechselwirkungen beteiligt. Das Verständnis der Bedeutung und Regulation von einzelnen *CYP3A*-Genen in der Pharmakologie und Physiologie ist unvollständig. Im Rahmen dieser Arbeit wurde die Evolution des *CYP3*-Genlokus über einen Zeitraum von 450 Millionen Jahren mittels genomischer Sequenzen von 16 Tierarten untersucht. Neue *CYP3*-Unterfamilien (*CYP3B*, *C* und *D*) entstanden mit erhöhter Evolutionsrate aus CYP3A-Vorstufen, die in rezenten Vertetern der Clupeocephala nachweisbar sind. Ausgeprägte funktionelle Unterschiede traten zwischen den *CYP3A* Orthologen von Säugern und Clupeocephala-Vertretern auf. Alle *CYP3A*-Gene der Amniota entwickelten sich aus zwei *CYP3A*-Urgenen. Mit der Entstehung von Säugern mit Plazenta (Eutheria) ging eines von ihnen verloren während das andere in eine neue genomische Umgebung transloziert wurde. Innerhalb der Primatendivergenz führten mehrere Genduplikationen, Deletionen, Pseudogenisierung und Genkonversionen zu einer raschen evolutionären Veränderung des *CYP3A*-Lokus. Die Entwicklung von *CYP3A* in catarrhinen Primaten (Altweltaffen, Menschenaffen und Menschen) unterschied sich wesentlich von Neuwelt-Primaten (z.B. Weißbüschelaffen) und Feuchtnasenaffen (z.B. Galago). In den Protein-codierenden Sequenzen von *CYP3A* sind Signaturen zweier früher Episoden von besonders starker positiver Selektion nachweisbar: (1) auf *CYP3A7* in der frühen hominoiden Evolution, die von einer Einschränkung der hepatischen Expression auf das Fötalstadium begleitet war, und (2) auf humanes *CYP3A4* nach dem letzten gemeinsamen Vorfahren von Schimpanse und Mensch. In Übereinstimmung mit diesen Befunden beeinflussen drei von vier positiv selektierten Aminosäuren, die in früheren biochemischen CYP3A-Studien untersucht wurden, die Aktivität und Regioselektivität. Es ist somit naheliegend, dass CYP3A7 und CYP3A4 katalytische Funktionen erworben haben können, die besonders wichtig für die Evolution von Menschenaffen und Menschen waren. Die Charakterisierung von *CYP3A*-Promotoren zeigte eine Anreicherung von ER6-Elementen bei Primaten. Die gestiegene Anzahl an ER6-Elementen kann für die ausgeprägte *CYP3A4*-Induzierbarkeit und Expressionsvariabilität im Menschen verantwortlich sein.

6. Abbreviations

BAC	bacterial artificial chromosome
BLAST	Basic Local Alignment and Search Tool
BLAT	BLAST-Like Alignment Tool
C/EBP	CCAAT-enhancer-binding proteins
CAR	constitutive androstane receptor
cDNA	complementary DNA
CLEM	Constitutive Liver Enhancer Module
CYP	Cytochrome P450
DHEAS	Dehydroepiandrosterone sulfate
DNA	Deoxyribonucleic acid
DR	direct repeat
ER	everted repeat
EST	expressed sequence tag
FXR	farnesoid X receptor
GR	glucocorticoid receptor
HNF4α	hepatocyte nuclear factor-4-a
HTG	High Throughput Genomic (sequence)
IR	inverted repeat
Kb	Kilo-base
LINE	long interspersed elements
LTR	likelihood ratio test
NWM	New World monkeys
ORF	open reading frame
OWM	Old World monkeys
PCR	polymerase chain reaction
PXR	pregnane X receptor
UTR	untranslated region
VDR	vitamin D receptor
WGS	Whole Genome Shotgun Sequence
XREM	Xenobiotic-Responsive Enhancer Module

7. References

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

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21:** 2104-2105.

Atkins, W.M. 2005. Non-Michaelis-Menten kinetics in cytochrome P450-catalyzed reactions. *Annu Rev Pharmacol Toxicol* **45:** 291-310.

Barber, D.S., A.J. McNally, N. Garcia-Reyero, and N.D. Denslow. 2007. Exposure to p,p'-DDE or dieldrin during the reproductive season alters hepatic CYP expression in largemouth bass (Micropterus salmoides). *Aquat Toxicol* **81:** 27-35.

Barwick, J.L., L.C. Quattrochi, A.S. Mills, C. Potenza, R.H. Tukey, and P.S. Guzelian. 1996. Trans-species gene transfer for analysis of glucocorticoid-inducible transcriptional activation of transiently expressed human CYP3A4 and rabbit CYP3A6 in primary cultures of adult rat and rabbit hepatocytes. *Mol Pharmacol* **50:** 10-16.

Bertilsson, G., A. Berkenstam, and P. Blomquist. 2001. Functionally conserved xenobiotic responsive enhancer in cytochrome P450 3A7. *Biochem Biophys Res Commun* **280:** 139-144.

Bieche, I., C. Narjoz, T. Asselah, S. Vacher, P. Marcellin, R. Lidereau, P. Beaune, and I. de Waziers. 2007. Reverse transcriptase-PCR quantification of mRNA levels from cytochrome (CYP)1, CYP2 and CYP3 families in 22 different human tissues. *Pharmacogenet Genomics* **17:** 731-742.

Bonfield, J.K., K. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res* **23:** 4992-4999.

Bork, R.W., T. Muto, P.H. Beaune, P.K. Srivastava, R.S. Lloyd, and F.P. Guengerich. 1989. Characterization of mRNA species related to human liver cytochrome P-450 nifedipine oxidase and the regulation of catalytic activity. *J Biol Chem* **264:** 910-919.

Brudno, M., C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13:** 721-731.

Burk, O., I. Koch, J. Raucy, E. Hustert, M. Eichelbaum, J. Brockmoller, U.M. Zanger, and L. Wojnowski. 2004. The induction of cytochrome P450 3A5 (CYP3A5) in the human liver and intestine is mediated by the xenobiotic sensors pregnane X receptor (PXR) and constitutively activated receptor (CAR). *J Biol Chem* **279:** 38379-38385.

Burk, O., H. Tegude, I. Koch, E. Hustert, R. Wolbold, H. Glaeser, K. Klein, M.F. Fromm, A.K. Nuessler, P. Neuhaus et al. 2002. Molecular mechanisms of polymorphic CYP3A7 expression in adult human liver and intestine. *J Biol Chem* **277:** 24280-24288.

Burk, O. and L. Wojnowski. 2004. Cytochrome P450 3A and their regulation. *Naunyn Schmiedebergs Arch Pharmacol* **369:** 105-124.

Busi, F. and T. Cresteil. 2005. CYP3A5 mRNA degradation by nonsense-mediated mRNA decay. *Mol Pharmacol* **68:** 808-815.

Carino, F.A., J.F. Koener, F.W. Plapp, Jr., and R. Feyereisen. 1994. Constitutive overexpression of the cytochrome P450 gene CYP6A1 in a house fly strain with metabolic resistance to insecticides. *Insect Biochem Mol Biol* **24:** 411-418.

Chen, H., A.G. Fantel, and M.R. Juchau. 2000. Catalysis of the 4-hydroxylation of retinoic acids by cyp3a7 in human fetal hepatic tissues. *Drug Metab Dispos* **28:** 1051-1057.

Claudianos, C., H. Ranson, R.M. Johnson, S. Biswas, M.A. Schuler, M.R. Berenbaum, R. Feyereisen, and J.G. Oakeshott. 2006. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol* **15:** 615-636.

Corley-Smith, G.E., H.T. Su, J.L. Wang-Buhler, H.P. Tseng, C.H. Hu, T. Hoang, W.G. Chung, and D.R. Buhler. 2006. CYP3C1, the first member of a new cytochrome P450 subfamily found in zebrafish (Danio rerio). *Biochem Biophys Res Commun* **340:** 1039-1046.

da Fonseca, R.R., A. Antunes, A. Melo, and M.J. Ramos. 2007. Structural divergence and adaptive evolution in mammalian cytochromes P450 2C. *Gene* **387:** 58-66.

Daly, A.K. 2006. Significance of the minor cytochrome P450 3A isoforms. *Clin Pharmacokinet* **45:** 13-31.

Danielson, P.B., R.J. MacIntyre, and J.C. Fogleman. 1997. Molecular cloning of a family of xenobiotic-inducible drosophilid cytochrome p450s: evidence for involvement in host-plant allelochemical resistance. *Proc Natl Acad Sci U S A* **94:** 10797-10802.

Davydov, D.R., J.R. Halpert, J.P. Renaud, and G. Hui Bon Hoa. 2003. Conformational heterogeneity of cytochrome P450 3A4 revealed by high pressure spectroscopy. *Biochem Biophys Res Commun* **312:** 121-130.

Dennison, J.B., P. Kulanthaivel, R.J. Barbuch, J.L. Renbarger, W.J. Ehlhardt, and S.D. Hall. 2006. Selective metabolism of vincristine in vitro by CYP3A5. *Drug Metab Dispos* **34:** 1317-1327.

Domanski, T.L., C. Finta, J.R. Halpert, and P.G. Zaphiropoulos. 2001a. cDNA cloning and initial characterization of CYP3A43, a novel human cytochrome P450. *Mol Pharmacol* **59:** 386-392.

Domanski, T.L. and J.R. Halpert. 2001. Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. *Curr Drug Metab* **2:** 117-137.

Domanski, T.L., Y.A. He, K.K. Khan, F. Roussel, Q. Wang, and J.R. Halpert. 2001b. Phenylalanine and tryptophan scanning mutagenesis of CYP3A4 substrate recognition site residues and effect on substrate oxidation and cooperativity. *Biochemistry* **40:** 10150-10160.

Evans, W.E. and M.V. Relling. 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286:** 487-491.

Feyereisen, R. 2006. Evolution of insect P450. *Biochem Soc Trans* **34:** 1252-1255.

Finta, C. and P.G. Zaphiropoulos. 2000. The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons. *Gene* **260:** 13-23.

Finta, C. and P.G. Zaphiropoulos. 2002. Intergenic mRNA molecules resulting from trans-splicing. *J Biol Chem* **277:** 5882-5890.

Frank, C., M.M. Gonzalez, C. Oinonen, T.W. Dunlop, and C. Carlberg. 2003. Characterization of DNA complexes formed by the nuclear receptor constitutive androstane receptor. *J Biol Chem* **278:** 43299-43310.

Frank, C., H. Makkonen, T.W. Dunlop, M. Matilainen, S. Vaisanen, and C. Carlberg. 2005. Identification of pregnane X receptor binding sites in the regulatory regions of genes involved in bile acid homeostasis. *J Mol Biol* **346:** 505-519.

Fraser, J.D., V. Martinez, R. Straney, and M.R. Briggs. 1998. DNA binding and transcription activation specificity of hepatocyte nuclear factor 4. *Nucleic Acids Res* **26:** 2702-2707.

Galetin, A., C. Brown, D. Hallifax, K. Ito, and J.B. Houston. 2004. Utility of recombinant enzyme kinetics in prediction of human clearance: impact of variability, CYP3A5, and CYP2C19 on CYP3A4 probe substrates. *Drug Metab Dispos* **32:** 1411-1420.

Gellner, K., R. Eiselt, E. Hustert, H. Arnold, I. Koch, M. Haberl, C.J. Deglmann, O. Burk, D. Buntefuss, S. Escher et al. 2001. Genomic organization of the human CYP3A locus: identification of a new, inducible CYP3A gene. *Pharmacogenetics* **11:** 111-121.

Givens, R.C., Y.S. Lin, A.L. Dowling, K.E. Thummel, J.K. Lamba, E.G. Schuetz, P.W. Stewart, and P.B. Watkins. 2003. CYP3A5 genotype predicts renal CYP3A activity and blood pressure in healthy adults. *J Appl Physiol* **95:** 1297-1300.

Gnerre, C., S. Blattler, M.R. Kaufmann, R. Looser, and U.A. Meyer. 2004. Regulation of CYP3A4 by the bile acid receptor FXR: evidence for functional binding sites in the CYP3A4 gene. *Pharmacogenetics* **14:** 635-645.

Goodwin, B., E. Hodgson, D.J. D'Costa, G.R. Robertson, and C. Liddle. 2002. Transcriptional regulation of the human CYP3A4 gene by the constitutive androstane receptor. *Mol Pharmacol* **62:** 359-365.

Goodwin, B., E. Hodgson, and C. Liddle. 1999. The orphan human pregnane X receptor mediates the transcriptional activation of CYP3A4 by rifampicin through a distal enhancer module. *Mol Pharmacol* **56:** 1329-1339.

Gotoh, O. 1992. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J Biol Chem* **267:** 83-90.

Graham, G.J. 1995. Tandem genes and clustered genes. *J Theor Biol* **175:** 71-87.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* **16:** 1664-1674.

Gu, X. and K. Vander Velden. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* **18:** 500-501.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696-704.

Guindon, S., F. Lethiec, P. Duroux, and O. Gascuel. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33:** W557-559.

Hakkola, J., O. Pelkonen, M. Pasanen, and H. Raunio. 1998. Xenobiotic-metabolizing cytochrome P450 enzymes in the human feto-placental unit: role in intrauterine toxicity. *Crit Rev Toxicol* **28:** 35-72.

Hara, H., Y. Yasunami, and T. Adachi. 2004. Loss of CYP3A7 gene induction by 1,25-dihydroxyvitamin D3 is caused by less binding of VDR to the proximal ER6 in CYP3A7 gene. *Biochem Biophys Res Commun* **321:** 909-915.

Harlow, G.R. and J.R. Halpert. 1997. Alanine-scanning mutagenesis of a putative substrate recognition site in human cytochrome P450 3A4. Role of residues 210 and 211 in flavonoid activation and substrate specificity. *J Biol Chem* **272:** 5396-5402.

He, Y.A., F. Roussel, and J.R. Halpert. 2003. Analysis of homotropic and heterotropic cooperativity of diazepam oxidation by CYP3A4 using site-directed mutagenesis and kinetic modeling. *Arch Biochem Biophys* **409:** 92-101.

Hinrichs, A.S., D. Karolchik, R. Baertsch, G.P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T.S. Furey, R.A. Harte, F. Hsu et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590-598.

Hirotsune, S., N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Takahashi, K. Yagami, A. Wynshaw-Boris, and A. Yoshiki. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423:** 91-96.

Huang, W., Y.S. Lin, D.J. McConn, 2nd, J.C. Calamia, R.A. Totah, N. Isoherranen, M. Glodowski, and K.E. Thummel. 2004. Evidence of significant contribution from CYP3A5 to hepatic drug metabolism. *Drug Metab Dispos* **32:** 1434-1445.

Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts et al. 2007. Ensembl 2007. *Nucleic Acids Res* **35:** D610-617.

Huelsenbeck, J.P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754-755.

Hukkanen, J., T. Vaisanen, A. Lassila, R. Piipari, S. Anttila, O. Pelkonen, H. Raunio, and J. Hakkola. 2003. Regulation of CYP3A5 by glucocorticoids and cigarette smoke in human lung-derived cells. *J Pharmacol Exp Ther* **304:** 745-752.

Hustert, E., M. Haberl, O. Burk, R. Wolbold, Y.Q. He, K. Klein, A.C. Nuessler, P. Neuhaus, J. Klattig, R. Eiselt et al. 2001. The genetic determinants of the CYP3A5 polymorphism. *Pharmacogenetics* **11:** 773-779.

Jeanmougin, F., J.D. Thompson, M. Gouy, D.G. Higgins, and T.J. Gibson. 1998. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* **23:** 403-405.

Jiang, G., L. Nepomuceno, K. Hopkins, and F.M. Sladek. 1995. Exclusive homodimerization of the orphan receptor hepatocyte nuclear factor 4 defines a new subclass of nuclear receptors. *Mol Cell Biol* **15:** 5131-5143.

Johnson, E.F. and C.D. Stout. 2005. Structural diversity of human xenobiotic-metabolizing cytochrome P450 monooxygenases. *Biochem Biophys Res Commun* **338:** 331-336.

Kamdem, L.K., F. Streit, U.M. Zanger, J. Brockmoller, M. Oellerich, V.W. Armstrong, and L. Wojnowski. 2005. Contribution of CYP3A5 to the in vitro hepatic clearance of tacrolimus. *Clin Chem* **51:** 1374-1381.

Kenworthy, K.E., S.E. Clarke, J. Andrews, and J.B. Houston. 2001. Multisite kinetic models for CYP3A4: simultaneous activation and inhibition of diazepam and testosterone metabolism. *Drug Metab Dispos* **29:** 1644-1651.

Khaitovich, P., W. Enard, M. Lachmann, and S. Paabo. 2006. Evolution of primate gene expression. *Nat Rev Genet* **7:** 693-702.

Khan, K.K. and J.R. Halpert. 2000. Structure-function analysis of human cytochrome P450 3A4 using 7-alkoxycoumarins as active-site probes. *Arch Biochem Biophys* **373:** 335-345.

Khan, K.K., Y.Q. He, T.L. Domanski, and J.R. Halpert. 2002. Midazolam oxidation by cytochrome P450 3A4 and active-site mutants: an evaluation of multiple binding sites and of the metabolic pathway that leads to enzyme inactivation. *Mol Pharmacol* **61:** 495-506.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29:** 170-179.

Kitada, M., T. Kamataki, K. Itahashi, T. Rikihisa, and Y. Kanakubo. 1987. P-450 HFLa, a form of cytochrome P-450 purified from human fetal livers, is the 16 alpha-hydroxylase of dehydroepiandrosterone 3-sulfate. *J Biol Chem* **262:** 13534-13537.

Kitada, M., T. Kamataki, K. Itahashi, T. Rikihisa, R. Kato, and Y. Kanakubo. 1985. Purification and properties of cytochrome P-450 from homogenates of human fetal livers. *Arch Biochem Biophys* **241:** 275-280.

Klingenberg, M. 1958. Pigments of rat liver microsomes. *Arch Biochem Biophys* **75:** 376-386.

Koch, I., R. Weil, R. Wolbold, J. Brockmoller, E. Hustert, O. Burk, A. Nuessler, P. Neuhaus, M. Eichelbaum, U. Zanger et al. 2002. Interindividual variability and tissue-specificity in the expression of cytochrome P450 3A mRNA. *Drug Metab Dispos* **30:** 1108-1114.

Koehler, S.C., N. Von Ahsen, C. Schlumbohm, A.R. Asif, U. Goedtel-Armbrust, M. Oellerich, L. Wojnowski, and V.W. Armstrong. 2006. Marmoset CYP3A21, a model for human CYP3A4: protein expression and functional characterization of the promoter. *Xenobiotica* **36:** 1210-1226.

Koley, A.P., J.T. Buters, R.C. Robinson, A. Markowitz, and F.K. Friedman. 1997. Differential mechanisms of cytochrome P450 inhibition and activation by alpha-naphthoflavone. *J Biol Chem* **272:** 3149-3152.

Korytko, P.J., F.W. Quimby, and J.G. Scott. 2000. Metabolism of phenanthrene by house fly CYP6D1 and dog liver cytochrome P450. *J Biochem Mol Toxicol* **14:** 20-25.

Korytko, P.J. and J.G. Scott. 1998. CYP6D1 protects thoracic ganglia of houseflies from the neurotoxic insecticide cypermethrin. *Arch Insect Biochem Physiol* **37:** 57-63.

Krasowski, M.D., K. Yasuda, L.R. Hagey, and E.G. Schuetz. 2005. Evolution of the pregnane x receptor: adaptation to cross-species differences in biliary bile salts. *Mol Endocrinol* **19:** 1720-1739.

Kriegs, J.O., G. Churakov, M. Kiefmann, U. Jordan, J. Brosius, and J. Schmitz. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol* **4:** e91.

Kuehl, P., J. Zhang, Y. Lin, J. Lamba, M. Assem, J. Schuetz, P.B. Watkins, A. Daly, S.A. Wrighton, S.D. Hall et al. 2001. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* **27:** 383-391.

Lacroix, D., M. Sonnier, A. Moncion, G. Cheron, and T. Cresteil. 1997. Expression of CYP3A in the human liver--evidence that the shift between CYP3A7 and CYP3A4 occurs immediately after birth. *Eur J Biochem* **247:** 625-634.

Lamba, J.K., Y.S. Lin, E.G. Schuetz, and K.E. Thummel. 2002. Genetic contribution to variable human CYP3A-mediated metabolism. *Adv Drug Deliv Rev* **54:** 1271-1294.

Laperriere, D., T.T. Wang, J.H. White, and S. Mader. 2007. Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. *BMC Genomics* **8:** 23.

Lee, A.J., A.H. Conney, and B.T. Zhu. 2003. Human cytochrome P450 3A7 has a distinct high catalytic activity for the 16alpha-hydroxylation of estrone but not 17beta-estradiol. *Cancer Res* **63:** 6532-6536.

Leeder, J.S., R. Gaedigk, K.A. Marcucci, A. Gaedigk, C.A. Vyhlidal, B.P. Schindel, and R.E. Pearce. 2005. Variability of CYP3A7 expression in human fetal liver. *J Pharmacol Exp Ther* **314:** 626-635.

Lemos, B., B.R. Bettencourt, C.D. Meiklejohn, and D.L. Hartl. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* **22:** 1345-1354.

Leonard, W.R. 2002. Food for thought. Dietary change was a driving force in human evolution. *Sci Am* **287:** 106-115.

Leowattana, W. 2004. DHEAS as a new diagnostic tool. *Clin Chim Acta* **341:** 1-15.

Li, X., J. Baudry, M.R. Berenbaum, and M.A. Schuler. 2004. Structural and functional divergence of insect CYP6B proteins: From specialist to generalist cytochrome P450. *Proc Natl Acad Sci U S A* **101:** 2939-2944.

Liu, F.J., X. Song, D. Yang, R. Deng, and B. Yan. 2008. The far and distal enhancers in the CYP3A4 gene co-ordinate the proximal promoter in responding similarly to the pregnane X receptor but differentially to hepatocyte nuclear factor-4alpha. *Biochem J* **409:** 243-250.

Magness, C.L., P.C. Fellin, M.J. Thomas, M.J. Korth, M.B. Agy, S.C. Proll, M. Fitzgibbon, C.A. Scherer, D.G. Miner, M.G. Katze et al. 2005. Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. *Genome Biol* **6:** R60.

Mangelsdorf, D.J., C. Thummel, M. Beato, P. Herrlich, G. Schutz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon et al. 1995. The nuclear receptor superfamily: the second decade. *Cell* **83:** 835-839.

Marill, J., T. Cresteil, M. Lanotte, and G.G. Chabot. 2000. Identification of human cytochrome P450s involved in the formation of all-trans-retinoic acid principal metabolites. *Mol Pharmacol* **58:** 1341-1348.

Martin, D. and E. Rybicki. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16:** 562-563.

Martinez-Jimenez, C.P., M.J. Gomez-Lechon, J.V. Castell, and R. Jover. 2005. Transcriptional regulation of the human hepatic CYP3A4: identification of a new distal enhancer region responsive to CCAAT/enhancer-binding protein beta isoforms (liver activating protein and liver inhibitory protein). *Mol Pharmacol* **67:** 2088-2101.

Martinez-Jimenez, C.P., R. Jover, M.T. Donato, J.V. Castell, and M.J. Gomez-Lechon. 2007. Transcriptional regulation and expression of CYP3A4 in hepatocytes. *Curr Drug Metab* **8:** 185-194.

Matsumura, K., T. Saito, Y. Takahashi, T. Ozeki, K. Kiyotani, M. Fujieda, H. Yamazaki, H. Kunitoh, and T. Kamataki. 2004. Identification of a novel polymorphic enhancer of the human CYP3A4 gene. *Mol Pharmacol* **65:** 326-334.

Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-1047.

McArthur, A.G., T. Hegelund, R.L. Cox, J.J. Stegeman, M. Liljenberg, U. Olsson, P. Sundberg, and M.C. Celander. 2003. Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. *J Mol Evol* **57:** 200-211.

Milton, K. 2003. The critical role played by animal source foods in human (Homo) evolution. *J Nutr* **133:** 3886S-3892S.

Nelson, D.R. 1998. Metazoan cytochrome P450 evolution. *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol* **121:** 15-22.

Nelson, D.R. 2003. Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. *Arch Biochem Biophys* **409:** 18-24.

Nelson, D.R., L. Koymans, T. Kamataki, J.J. Stegeman, R. Feyereisen, D.J. Waxman, M.R. Waterman, O. Gotoh, M.J. Coon, R.W. Estabrook et al. 1996. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6:** 1-42.

Nelson, D.R., D.C. Zeldin, S.M. Hoffman, L.J. Maltais, H.M. Wain, and D.W. Nebert. 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* **14:** 1-18.

Nielsen, R. and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929-936.

Nuzhdin, S.V., M.L. Wayne, K.L. Harmon, and L.M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol Biol Evol* **21:** 1308-1317.

Ohmori, S., H. Nakasa, K. Asanome, Y. Kurose, I. Ishii, M. Hosokawa, and M. Kitada. 1998. Differential catalytic properties in metabolism of endogenous and exogenous substrates among CYP3A enzymes expressed in COS-7 cells. *Biochim Biophys Acta* **1380:** 297-304.

Olefsky, J.M. 2001. Nuclear receptor minireview series. *J Biol Chem* **276:** 36863-36864.

Olson, M.V. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64:** 18-23.

Ozdemir, V., W. Kalow, B.K. Tang, A.D. Paterson, S.E. Walker, L. Endrenyi, and A.D. Kashuba. 2000. Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics* **10:** 373-388.

Paine, M.F., H.L. Hart, S.S. Ludington, R.L. Haining, A.E. Rettie, and D.C. Zeldin. 2006. The human intestinal cytochrome P450 "pie". *Drug Metab Dispos* **34:** 880-886.

Pal, C., B. Papp, and M.J. Lercher. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7:** 337-348.

Park, H., S. Lee, and J. Suh. 2005. Structural and dynamical basis of broad substrate specificity, catalytic mechanism, and inhibition of cytochrome P450 3A4. *J Am Chem Soc* **127:** 13634-13642.

Plant, N. 2007. The human cytochrome P450 sub-family: transcriptional regulation, inter-individual variation and interaction networks. *Biochim Biophys Acta* **1770:** 478-488.

Pond, S.L., S.D. Frost, and S.V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21:** 676-679.

Posada, D. and K.A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98:** 13757-13762.

Posada, D. and K.A. Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* **54:** 396-402.

Robinson-Rechavi, M. and D. Huchon. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics* **16:** 296-297.

Rodriguez-Antona, C., M. Axelson, C. Otter, A. Rane, and M. Ingelman-Sundberg. 2005. A novel polymorphic cytochrome P450 formed by splicing of CYP3A7 and the pseudogene CYP3AP1. *J Biol Chem* **280:** 28324-28331.

Rodriguez-Antona, C., R. Bort, R. Jover, N. Tindberg, M. Ingelman-Sundberg, M.J. Gomez-Lechon, and J.V. Castell. 2003. Transcriptional regulation of human CYP3A4 basal expression by CCAAT enhancer-binding protein alpha and hepatocyte nuclear factor-3 gamma. *Mol Pharmacol* **63:** 1180-1189.

Ronquist, F. and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Saito, T., Y. Takahashi, H. Hashimoto, and T. Kamataki. 2001. Novel transcriptional regulation of the human CYP3A7 gene by Sp1 and Sp3 through nuclear factor kappa B-like element. *J Biol Chem* **276:** 38010-38022.

Sandelin, A. and W.W. Wasserman. 2005. Prediction of nuclear hormone receptor response elements. *Mol Endocrinol* **19:** 595-606.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol* **6:** 526-538.

Schirmer, M., M.R. Toliat, M. Haberl, A. Suk, L.K. Kamdem, K. Klein, J. Brockmoller, P. Nurnberg, U.M. Zanger, and L. Wojnowski. 2006. Genetic signature consistent with selection against the CYP3A4*1B allele in non-African populations. *Pharmacogenet Genomics* **16:** 59-71.

Schleinkofer, K., Sudarko, P.J. Winn, S.K. Ludemann, and R.C. Wade. 2005. Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? *EMBO Rep* **6:** 584-589.

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

Schuetz, J.D., S. Kauma, and P.S. Guzelian. 1993. Identification of the fetal liver cytochrome CYP3A7 in human endometrium and placenta. *J Clin Invest* **92:** 1018-1024.

Schuetz, J.D., E.G. Schuetz, J.V. Thottassery, P.S. Guzelian, S. Strom, and D. Sun. 1996. Identification of a novel dexamethasone responsive enhancer in the human CYP3A5 gene and its activation in human and rat liver cells. *Mol Pharmacol* **49:** 63-72.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-586.

Shen, R.F. and H.H. Tai. 1998. Thromboxanes: synthase and receptors. *J Biomed Sci* **5:** 153-172.

Shimada, T., H. Yamazaki, M. Mimura, Y. Inui, and F.P. Guengerich. 1994. Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: studies with liver microsomes of 30 Japanese and 30 Caucasians. *J Pharmacol Exp Ther* **270:** 414-423.

Shimada, T., H. Yamazaki, M. Mimura, N. Wakamiya, Y.F. Ueng, F.P. Guengerich, and Y. Inui. 1996. Characterization of microsomal cytochrome P450 enzymes involved in the oxidation of xenobiotic chemicals in human fetal liver and adult lungs. *Drug Metab Dispos* **24:** 515-522.

Sim, S.C., R.J. Edwards, A.R. Boobis, and M. Ingelman-Sundberg. 2005. CYP3A7 protein expression is high in a fraction of adult human livers and partially associated with the CYP3A7*1C allele. *Pharmacogenet Genomics* **15:** 625-631.

Stajich, J.E., D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, H. Lapp et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12:** 1611-1618.

Staudinger, J.L., B. Goodwin, S.A. Jones, D. Hawkins-Brown, K.I. MacKenzie, A. LaTour, Y. Liu, C.D. Klaassen, K.K. Brown, J. Reinhard et al. 2001. The nuclear receptor PXR is a lithocholic acid sensor that protects against liver toxicity. *Proc Natl Acad Sci U S A* **98:** 3369-3374.

Stevens, J.L., M.J. Snyder, J.F. Koener, and R. Feyereisen. 2000. Inducible P450s of the CYP9 family from larval Manduca sexta midgut. *Insect Biochem Mol Biol* **30:** 559-568.

Takiguchi, T., M. Tomita, N. Matsunaga, H. Nakagawa, S. Koyanagi, and S. Ohdo. 2007. Molecular basis for rhythmic expression of CYP3A4 in serum-shocked HepG2 cells. *Pharmacogenet Genomics* **17:** 1047-1056.

Tegude, H., A. Schnabel, U.M. Zanger, K. Klein, M. Eichelbaum, and O. Burk. 2007. Molecular mechanism of basal CYP3A4 regulation by hepatocyte nuclear factor 4alpha: evidence for direct regulation in the intestine. *Drug Metab Dispos* **35:** 946-954.

Thomas, J.H. 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genet* **3:** e67.

Thompson, E.E., H. Kuttab-Boulos, D. Witonsky, L. Yang, B.A. Roe, and A. Di Rienzo. 2004. CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet* **75:** 1059-1069.

Thompson, P.D., P.W. Jurutka, G.K. Whitfield, S.M. Myskowski, K.R. Eichhorst, C.E. Dominguez, C.A. Haussler, and M.R. Haussler. 2002. Liganded VDR induces CYP3A4 in small intestinal and colon cancer cells via DR3 and ER6 vitamin D responsive elements. *Biochem Biophys Res Commun* **299:** 730-738.

Thummel, K.E., C. Brimer, K. Yasuda, J. Thottassery, T. Senn, Y. Lin, H. Ishizuka, E. Kharasch, J. Schuetz, and E. Schuetz. 2001. Transcriptional control of intestinal cytochrome P-4503A by 1alpha,25-dihydroxy vitamin D3. *Mol Pharmacol* **60:** 1399-1406.

Tijet, N., C. Helvig, and R. Feyereisen. 2001. The cytochrome P450 gene superfamily in Drosophila melanogaster: annotation, intron-exon organization and phylogeny. *Gene* **262:** 189-198.

Tirona, R.G., W. Lee, B.F. Leake, L.B. Lan, C.B. Cline, V. Lamba, F. Parviz, S.A. Duncan, Y. Inoue, F.J. Gonzalez et al. 2003. The orphan nuclear receptor HNF4alpha determines PXR- and CAR-mediated xenobiotic induction of CYP3A4. *Nat Med* **9:** 220-224.

Torimoto, N., I. Ishii, K.I. Toyama, M. Hata, K. Tanaka, H. Shimomura, H. Nakamura, N. Ariyoshi, S. Ohmori, and M. Kitada. 2006. Helices F-G are important for the substrate specificities of CYP3A7. *Drug Metab Dispos*.

Vaccaro, E., A. Salvetti, R.D. Carratore, S. Nencioni, V. Longo, and P.G. Gervasi. 2007. Cloning, tissue expression, and inducibility of CYP 3A79 from sea bass (Dicentrarchus labrax). *J Biochem Mol Toxicol* **21:** 32-40.

Verschure, P.J. 2004. Positioning the genome within the nucleus. *Biol Cell* **96:** 569-577.

Verslycke, T., J.V. Goldstone, and J.J. Stegeman. 2006. Isolation and phylogeny of novel cytochrome P450 genes from tunicates (Ciona spp.): a CYP3 line in early deuterostomes? *Mol Phylogenet Evol* **40:** 760-771.

Wang, H., R. Dick, H. Yin, E. Licad-Coles, D.L. Kroetz, G. Szklarz, G. Harlow, J.R. Halpert, and M.A. Correia. 1998. Structure-function relationships of human liver cytochromes P450 3A: aflatoxin B1 metabolism as a probe. *Biochemistry* **37:** 12536-12545.

Wen, B., C.E. Doneanu, J.N. Lampe, A.G. Roberts, W.M. Atkins, and S.D. Nelson. 2005. Probing the CYP3A4 active site by cysteine scanning mutagenesis and photoaffinity labeling. *Arch Biochem Biophys* **444:** 100-111.

Werck-Reichhart, D. and R. Feyereisen. 2000. Cytochromes P450: a success story. *Genome Biol* **1:** REVIEWS3003.

Wilkinson, G.R. 2005. Drug metabolism and variability among patients in drug response. *N Engl J Med* **352:** 2211-2221.

Williams, E.J. and L.D. Hurst. 2000. The proteins of linked genes evolve at similar rates. *Nature* **407:** 900-903.

Williams, E.T., A.S. Rodin, and H.W. Strobel. 2004a. Defining relationships between the known members of the cytochrome P450 3A subfamily, including five putative chimpanzee members. *Mol Phylogenet Evol* **33:** 300-308.

Williams, E.T., K.R. Schouest, M. Leyk, and H.W. Strobel. 2007. The chimpanzee cytochrome P450 3A subfamily: Is our closest related species really that similar?. Comp Biochem Physiol Part D Genomics Proteomics 2: 91-100.

Williams, J.A., B.J. Ring, V.E. Cantrell, D.R. Jones, J. Eckstein, K. Ruterbories, M.A. Hamman, S.D. Hall, and S.A. Wrighton. 2002. Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7. *Drug Metab Dispos* **30:** 883-891.

Williams, P.A., J. Cosme, D.M. Vinkovic, A. Ward, H.C. Angove, P.J. Day, C. Vonrhein, I.J. Tickle, and H. Jhoti. 2004b. Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **305:** 683-686.

Wojnowski, L. 2004. Genetics of the variable expression of CYP3A in humans. *Ther Drug Monit* **26:** 192-199.

Wojnowski, L. and L.K. Kamdem. 2006. Clinical implications of CYP3A polymorphisms. *Expert Opin Drug Metab Toxicol* **2:** 171-182.

Xie, W., J.L. Barwick, C.M. Simon, A.M. Pierce, S. Safe, B. Blumberg, P.S. Guzelian, and R.M. Evans. 2000. Reciprocal activation of xenobiotic response genes by nuclear receptors SXR/PXR and CAR. *Genes Dev* **14:** 3014-3023.

Xue, L., V.G. Zgoda, B. Arison, and M. Almira Correia. 2003. Structure-function relationships of rat liver CYP3A9 to its human liver orthologs: site-directed active site mutagenesis to a progesterone dihydroxylase. *Arch Biochem Biophys* **409:** 113-126.

Yamaori, S., H. Yamazaki, A. Suzuki, A. Yamada, H. Tani, T. Kamidate, K. Fujita, and T. Kamataki. 2003. Effects of cytochrome b(5) on drug oxidation activities of human cytochrome P450 (CYP) 3As: similarity of CYP3A5 with CYP3A4 but not CYP3A7. *Biochem Pharmacol* **66:** 2333-2340.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15:** 568-573.

Yang, Z. and J.P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15:** 496-503.

Yang, Z., R. Nielsen, N. Goldman, and A.M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431-449.

Zhang, Z.D., P. Cayting, G. Weinstock, and M. Gerstein. 2008. Analysis of nuclear receptor pseudogenes in vertebrates: how the silent tell their stories. *Mol Biol Evol* **25:** 131-143.

8. Appendix

Table 8.1. Source of genomic sequences

Gene	sequence ID	Start	End	Length (bp)	strand
***CYP3A* genomic sequences**					
Calja3A21	EF600758.1*	98255	122418	24163	+
Calja3A5	EF600758.1*	49625	88132	38507	+
Calja3A90	EF600758.1*	4192	38492	34300	+
Homsa3A4	chr7	99141059	99170652	29593	-
Homsa3A43	chr7	99263675	99301560	37885	+
Homsa3A5	chr7	99083864	99115455	31591	-
Homsa3A7	chr7	99193692	99219640	25948	-
Mulma3A4	chr3	47026931	47056683	29753	-
Mulma3A43	chr3	47099079	47140688	41610	+
Mulma3A5	chr3	46853057	46879943	26887	-
Mulma3A7	chr3	46903538	46929928	26391	-
Otoga3A91	EF600756*	112978	127174	14196	+
Otoga3A91	EF600756*	1	22486	22485	+
Otoga3A92	EF600756*	28456	53063	24607	-
Pantr3A4	chr7	99609109	99635999	26878	-
Pantr3A43	chr7	99674258	99713284	39027	+
Pantr3A5	chr7	99423698	99456115	32418	-
Pantr3A67	chr7	99485118	99514183	29066	-
Pantr3A7	chr7	99555524	99582881	27358	-
Papan3A4	AC141417.16*	108891	139083	30192	-
Papan3A7	AC141417.16*	9579	33459	23880	-
***CYP3A* promoter sequences (RE conservation analysis)**					
Calja3A21	EF600758.1*	136202	122419	13783	+
Calja3A5	EF600758.1*	95902	88133	6960	+
Calja3A90	EF600758.1*	45788	38943	6847	+
Homsa3A4	chr7	99219641	99232640	13000	+

Homsa3A43	chr7	99250675	99263674	13000	-
Homsa3A5	chr7	99115456	99128455	13000	-
Homsa3A7	chr7	99170653	99184065	13413	-
Mulma3A4	chr3	47056684	47069929	13246	-
Mulma3A43	chr3	47086315	47099078	12764	+
Mulma3A5	chr3	46879944	46890757	10751	-
Mulma3A7	chr3	46929929	46945865	15937	-
Otoga3A91	EF600756*	101926	112977	11052	+
Otoga3A92	EF600756*	53064	68205	15141	-
Pantr3A4	chr7	99636000	99649026	13026	-
Pantr3A43	chr7	99661445	99674257	12974	+
Pantr3A5	chr7	99456116	99468371	12255	-
Pantr3A67	AC145951.2*	42413	52145	9733	-
Pantr3A7	chr7	99582882	99597345	14454	-
Papan3A4	AC141417.16*	139084	150801	12861	-
Papan3A7	AC141417.16*	33632	49653	16022	-

***CYP3A* promoter sequences (RE density analysis)**

Calja3A21	EF600758.1*	109420	122419	13000	+
Calja3A5	EF600758.1*	75134	88133	13000	+
Calja3A90	EF600758.1*	25944	38943	13000	+
Homsa3A4	chr7	99219641	99232640	13000	+
Homsa3A43	chr7	99250675	99263674	13000	-
Homsa3A5	chr7	99115456	99128455	13000	-
Homsa3A7	chr7	99170653	99183652	13000	-
Mulma3A4	chr3	47056684	47069683	13000	-
Mulma3A43	chr3	47086079	47099078	13000	+
Mulma3A5	chr3	46879944	46892943	13000	-
Mulma3A7	chr3	46929929	46942928	13000	-
Otoga3A91	EF600756*	99978	112977	13000	+
Otoga3A92	EF600756*	53064	66063	13000	-
Pantr3A4	chr7	99636000	99648999	13000	-
Pantr3A43	chr7	99661258	99674257	13000	+
Pantr3A5	chr7	99456116	99469115	13000	-

Pantr3A67	AC145951.2*	42413	55412	13000	-
Pantr3A7	chr7	99582882	99595881	13000	-
Papan3A4	AC141417.16*	139084	152083	13000	-
Papan3A7	AC141417.16*	33632	46631	13000	-

*GenBank accession number. Homsa (*Homo sapiens*), Pantr (*Pan troglodytes*), Macmu (*Macaca mulatta*), Papan (*Papio anubis*), Calja (*Callithrix jacchus*), Otoga (*Otolemur garnettii*).

Table 8.2. Primers for primate *CYP3A* cDNA amplification

Primer	**Sequence (5'- 3')**	**Annealing temperature**	**Binding position**	**Amplification targets***
Primers for initial PCR amplification				
3A4e1F	AAAGAGCAACACAGAGCTG	54	Exon1	CYP3A4 (Pt, Pp)
3A4e13R	GTCCTTAGGAAAATTCAGG	54	Exon13	CYP3A4 (Pt, Pp)
3A5e1F	AAACAGCAGCACTCAGCTA	54	Exon1	CYP3A5 (Pt, Pp)
3A5e13R	AGTCCTTAGAATAACTCATT	54	Exon13	CYP3A5 (Pt, Pp)
3A67e1F	GCAGCACGCTGCTGAAAA	54	Exon1	CYP3A67 (Pt, Pp)
3A67e13R	CTTCATTTCAGGGTTCTATTTAT	54	Exon13	CYP3A67 (Pt, Pp)
M3A7e1F	CCTGGCTGTCAGCCTGATAC	56	Exon1	CYP3A7 (Mm)
M3A7e13R	CATCCCTTGACTCAGCCGTT	56	Exon13	CYP3A7 (Mm)
3A43e1F	CCCAGCAAAGAGCAGCACAC	54	Exon1	CYP3A43 (Pt)
3A43e13R	GGGATACAGCTTTCTTGAAC	54	Exon13	CYP3A43 (Pt)
CJ5F	GAAGACTCGGAGGAGAGAGATAA	57	Exon1	CYP3A5a/b (Cj)
CJ5R1	GCACAGCTTTCTTCAAGAGCA	57	Exon13	CYP3A5a (Cj)
CJ5R2	CTTTCTTCAAGAGCAAAGCAGT	57	Exon13	CYP3A5b (Cj)
Primers for nested PCR amplification				
3A4e1F2	TCAGAGGAGAGAGATAAGT	54	Exon1	5'-CYP3A4 (Pt, Pp)
3A5e1F2	TCACAGAAGACAGTTGAAG	54	Exon1	5'-CYP3A5 (Pt, Pp)
3A-1088R	GTTTCATTCACCACCATGTC	54	Exon7	5'-CYP3A4/5 (Pt, Pp)
3A-531F	GGCCTACAGCATGGATGTG	54	Exon10	3'-CYP3A4/5 (Pt, Pp)
3A4e13R2	TGCCATCCCTTGACTCA	54	Exon13	3'-CYP3A4 (Pt, Pp)
3A5e13R2	TCTCCATCTCTTGAATCC	54	Exon13	3'-CYP3A5 (Pt, Pp)
3A67e1F2	TCATCCCAAACTTGGCCG	54	Exon1	5'/3'-CYP3A67 (Pt, Pp)
3A67e13R2	AGCTTTCTTGAAGAGCAAAC	54	Exon13	5'/3'-CYP3A67 (Pt, Pp)
M3A7e1F	CCTGGCTGTCAGCCTGATAC	56	Exon1	5'-CYP3A7 (Mm)
M3A7e7R	GTAACTTTTCTTGGAAACACAGTG	56	Exon7	5'-CYP3A7 (Mm)
M3A7e7F	GAAGCTTTTAAGATTCAATTCATTA	56	Exon7	3'-CYP3A7 (Mm)
M3A7e13R	CATCCCTTGACTCAGCCGTT	56	Exon13	3'-CYP3A7 (Mm)
3A43e1F	CCCAGCAAAGAGCAGCACAC	55	Exon1	5'-CYP3A43 (Pp, Mm)
3A43e10R	ACTGCGTCAATCTCCTCCTG	55	Exon10	5'-CYP3A43 (Pp, Mm)
3A43e9F	CCGAGTAGATTTCTTTCAACAG	55	Exon9	3'-CYP3A43 (Pp, Mm)
3A43e13R2	GTGGAAGTCCTTAGGGAAAGTCAG	55	Exon13	3'-CYP3A43 (Pp, Mm)

* Pt (Pan troglodytes), Mm (Macaca mulatta), Pp (Pongo pygmaeus), Cj (Callithrix jacchus).

Table 8.3. Genomic coordinates of *CYP3A* Homologous Regions and other non-amniota *CYP3* loci.

	Genome		
Species	**assembly**	**Chromosome**	**Chromosome location**
CYP3HR1			
human	(hg18)	chr7	3307606~~4889861
chimpanzee	(panTro2)	chr7	3254896~~4846438
rhesus	(rheMac2)	chr3	42751381~~42877840
dog	(canFam2)	chr16	16145183~~15855321
horse	(Equus1.0)	(Genbank No.DS178473)	1732031~~2657694
mouse	(mm8)	chr5	142109261~~142803551
rat	(rn4)	chr12	12802275~~12485342
opossum	(monDom4)	chr6	79168894~~81492900
chicken	(galGal3)	chr14	3540926~~4044674
Zebrafish	(danRer4)	chr3	48232991~~48149500
fugu	(fr1)	scaffold_261	88284083~~88259414
stickleback	(BRAOAD1)	Group IX	4892313~~4851298
platypus		Contig540	37820~~453179
CYP3HR2			
human	(hg18)	chr7	98874632~~99411623
chimpan	(panTro2)	chr7	99205874~~99824073
rhesus	(rheMac2)	chr3	46645420~~47243091
dog	(canFam2)	chr16	13148149~~12635853
horse	(Equus1.0)	(Genbank No.DS178473)	255511~~1125573
mouse	(mm8)	chr5	145420751~~146774650
mouse	(mm8)	chr5	138021702~~138220018
rat	(rn4)	Chr12	9743026~~8981738
rat	(rn4)	Chr12	16825703~~17498949
CYP3B			
tetraodon	(tetNeig1)	Chr17	10597497~~10640828
fugu	(fr1)	scaffold_111	47053509~~47010768
stickleback	(BROAD1)	Group IV	14002463~~13932184
medaka	(HdrR)	chr15	14881820~~15022360
Pufferfish *CYP3A*			
tetraodon	(tetNeig1)	chr9	5117899~~5257374
fugu	(fr1)	scaffold_128	52453029~~52594931

frog *CYP3A*			
frog	(xenTro2)	scaffold_320	90781~~823355
frog	(xenTro2)	scaffold_76	34838~~179108

stickleback *CYP3A48*			
stickleback	(BROAD1)	groupXII	12750847~~12818251

zebrafish *CYP3A65*			
zebrafish	(danRer4)	chr1	66296045~~66403567

Table 8.4. Potential gene conversion events among *CYP3* genes detected by GENECONV.

Species	Genes involved	p value	Begin	Length	No. of differences	Exons involved
Protein coding region						
human*	*CYP3A4;CYP3A7*	0	99204361; 99151771[1]	1098;1096	42	6
chimpanzee*	*CYP3A4;CYP3A7*	0	99620616; 99566235[2]	1098;1096	45	6
orangutan*	*CYP3A4;CYP3A7*	0	N/A	1097;1100	46	6
orangutan*	*CYP3A67;CYP3A7*	0	N/A	958;996	9	3
marmoset*	*CYP3A90;CYP3A5*	0	38721;87908[3]	15979;12271	115	2-4
marmoset*	*CYP3A90;CYP3A5*	0.00022	14765;60480[3]	541;536	7	8
marmoset*	*CYP3A90;CYP3A5*	0	13416;59112[3]	2361;2355	26	9
marmoset*	*CYP3A90;CYP3A5*	0	6839;52301[3]	2670;2698	68	12-13
rat	*CYP3A1*[$]*;CYP3A2*	0.0004	361;348[4]	498; 498	0	4-8
dog	*CYP3A12;CYP3A98*	0.0001	922;921[4]	648;648	0	12-13
pig	*CYP3A39;CYP3A22*	0.0032	999;1007[4]	573;573	5	12-13
medaka	*CYP3B3;CYP3B4*	0.0094	N/A	147;147	0	12-13
Promoter region						
human	*CYP3A4;CYP3A7*	0.01464	99221245;99172259[1]	273;273	5	
rhesus	*CYP3A4;CYP3A7*	0.00714	47062130;46935392[5]	109;109	0	
rhesus	*CYP3A4;CYP3A7*	0.00001	47058449;46931686[5]	825;829	23	
rhesus	*CYP3A4;CYP3A7*	0.03303	47057668;46930906[5]	142;142	0	
baboon	*CYP3A4;CYP3A7*	0.00946	139130;33678[6]	307;309	6	

* These gene conversion events were detected by analysis of genomic sequences alignments. p values are global permutation values. No. of differences indicates the number of sequence differences which arose between the converted regions of the involved two sequences. Exon 6 conversions detected in the human, chimpanzee, and orangutan sequences reflect a single conversion event in a common ancestor of these species (see Fig. 2). [1]Begin of the conversion numbered using human chromosome 7 sequence (hg18); [2]Begin of the conversion numbered using chimpanzee chromosome 7 sequence (panTro2); [3]Begin of the conversion numbered using BAC sequence (Genbank accession No. EF600758); [4]Begin of the conversion in rat, dog and pig *CYP3A* numbered according to sequences: *CYP3A1* (NM_173144); *CYP3A2* (U09742); *CYP3A12* (NM_001003340); *CYP3A98* (XM_536868); *CYP3A39* (NM_214422); *CYP3A22* (AB006010). [5]Begin of the conversion numbered using rhesus chromosome 3 sequence (macMul2); [6]Begin of the conversion numbered using baboon BAC sequence (AC141417.16); [$]The rat *CYP3A1* sequence was poorly sequenced and may have some errors. Since the orangutan *CYP3A* genomic sequences and medaka *CYP3B* cDNA sequences are not available in public databases, the sequence sources are not provided.

Table 8.5. Genbank *CYP3C* and *CYP3D* EST sequences in Clupeocephala species different from fugu, tetraodon, and zebrafish. Genbank database was accessed in March 2007.

Species	*CYP3B* EST sequences (NCBI GI Number)
Hippoglossus hippoglossus	90597214
Lithognathus mormyrus	120472677,120472676, 120474737
Astatotilapia burtoni	89326979,120470950,89324926
Fundulus heteroclitus	68256846,66935956,68257408,68257563,66268628, 66935537,68256845
Species	*CYP3C* EST sequences (NCBI GI Number)
Pimephales promelas	73731960,73437962,
Misgurnus anguillicaudatus	64689867,64684549,64678463,64700279,64662492, 64654347

8.6. Marmoset *CYP3A* processed pseudogene sequence

>Marmoset *CYP3A* processed pseudogene

tatggaactaattcacatgggctttttaagaagcttggaattctgggacccacacctctgccctttttgggaactgttttatcctaccgaca
ggacttttggaagtttgacatggaatgttataaaaagtatggagaagtgtgggggatttatgatggtcgacagcctgtgctggctatcg
cagatcccaacataatcaaaacagtgctagtgaaagaatgttattctgtcttcgctaaccggaggtctttcggtccagtgggatttatga
aaagtgccatctctatagctacggatgataaatggaagaaaatatgatcattgctgtctccaatcttcaccagtggaaaactcaaggag
tccctatctttgcccagtatggagaggtgttggtgaaaaacctgaggcgggaagcagagaaaggcaaggaatcaacatgaaagac
atctttggggcctacagcatggatgtgatcactgacacgtcatttggagtgaacaacaactctctcaacaatccacaagacccctttgt
ggaaagcaccaagaagcttttaagatttgatgtttcagatccattctttctctcaataacaatctttccattccttaccccaattcttgaagc
attaaatatttctgtgtttccaagagattctacaagtttttaagaaaatctataaaaaggataaaagaaagtcgtctcaaagatacacata
agcaccgagtggatttccttcagctgatgattgactcccagaattcgaaagaaactgagtcccacaaagctctgtctgatctggagct
catggcccaatcaattatcttcatttttactggctatgaaaccaccagcagtagtcttttcttcgttatgtatgaactggccactcaccctga
tgtccagcagaaaccgcaggaggaaattgatgcagttttacccaacaaggcaccagccacctatgatactgtgctacagatggagt
atcttgacatggtggtgaatgaaacactcagattattcccacttgctatgagacttgagaggatctgcaaaaaagatgttgaaatcaatg
ggatgttgattcccaaaggggtagtggtgatgaatccaagctatgctcttcactatgacccaaagtactggacagagactgagaagtt
cctccctgaaaggttcagtaagaataacaaggacaacacagatagcacatatacacatcctttagaactggacccagaaactgcatt
ggcatgaggtttgctctcatgaacatggtacttgctctaatcagagtccttcagaacttgtccttcaaaccttgtaaagaaatacagatcc
ccctgaaattatgcttaggaggacttcttcaaacagaaaaacctattgttctaaaggttgagtcaagggatgggactgtaagtggaacc

Table 8.7. Annotation of nuclear receptor binding sites (by NHR-scan) in human *CYP3A* promoters

Site type	Sequence	Position*	Probability	Repeat	Repeat type
Human *CYP3A4*					
ER6&	TGAACTCAAAGGAGGTCA	167	5.7449		
ER6	GGAGCTCACCTCTGTTCA	590	1.416		
ER6	GGACCTTTGAAGGGTTCA	933	2.6568		
DR4	TGACCTGCAGTGACCA	1107	3.7538		
IR1	GGGTCAGGGAGCT	1288	1.9077		
DR1	TGAACTTTGATCC	3046	2.513		
ER6	TGAACTCTAGCCTGGGCA	3495	1.7276	AluJo	SINE/Alu
DR4	GGATCACTTGAGGCCA	3672	1.3381	AluJo	SINE/Alu
ER6	TGAAATCCCAGAACTTCA	3706	1.5718	AluJo	SINE/Alu
DR4	AGGCCAATTCTGGTCA	4244	0.837		
DR4&	TGTCCTGTGTTGACCC	7602	3.0487	Charlie22a	DNA/MER1
ER6&	TGAAATCATGTCGGTTCA	7673	2.6391		
DR3&	TGAACTTGCTGACCC	7717	4.0221		
DR4&	TAAACTGAGATGATCT	7741	1.1672		
ER6&	TGTCCCAATTAAAGGTCA	7773	2.492	Charlie4z	DNA/MER1
DR2	GGCTCAGTGGTTCA	8711	0.8541	AluJb	SINE/Alu
DR1	AGGGCAAAGGACA	8800	1.9432		
ER6	TCACCTTAAAATGGTTAA	9244	1.3563	L1MB5	LINE/L1
ER4	TGAGCTCAGGAGTTCA	10316	3.0522	AluSx	SINE/Alu
DR1	AGATCACAGGCCA	10595	1.5475		
ER1&	TGAACTTAGCTCA	11148	3.5884		
ER6&	TGAACTTCCTGAAGTTCA	11352	6.0684		
ER6	TGACCTGGAACCAATCCA	11959	1.9123	L1PA7	LINE/L1
DR3	AGCTCAGTAAGGCCA	12277	0.9382	L1P3	LINE/L1
ER6	TGAACTGGGTGGAGCTCA	12299	4.358	L1P3	LINE/L1
DR1	TGACCTGGGACCC	12409	2.099	L1P3	LINE/L1
Human *CYP3A7*					
ER6	TTAACTCAATGGAGGTCA	166	4.4077		
DR4	TGACCTGCAGTGACCA	1109	3.7538		
IR1	GGGTCACAGGGCT	1290	2.1291		
DR1	TGACCTGTGAAAT	3288	1.8792		
DR4	GGATCACTTGAGGCCA	3680	1.3381	AluJo	SINE/Alu
DR4	AGGCCAATTCTGGTCA	4243	0.9954		
ER8	TGAGCTGCATTGCTAGTTCC	4871	1.6338	MIR	SINE/MIR
DR0	AGGGCAAGGGCA	5000	1.3123		
DR2	TGACCTCATGATCT	6568	2.606	AluSg	SINE/Alu

IR1	AGTGCAATGGCCT	6762	1.0149	AluSg	SINE/Alu
DR4	TGCCCTATGCTGACCC	7599	3.5242	Charlie22a	DNA/MER1
ER6	TGAAATCATGTCAGTTCA	7670	4.1051		
DR3	TGAACTTGCTGACCC	7714	4.0221		
ER6	TGTCCCAACTAAAGTTCA	7770	2.3256		
DR2	AGGCCACAAGTTCA	8668	1.7172	AluJb	SINE/Alu
DR4	TGCCCTGGCCAGAACT	10323	0.9361	L1PA5	LINE/L1
IR1	AGGTCCCAGACCT	13099	0.9807		
Human *CYP3A5*					
ER6	TGAACTCAAAAGAGGTCA	125	5.8342		
ER1	TGAGCTGAGATCA	2059	1.5125	AluSx	SINE/Alu
DR2	AGGTCAGGAGTTCA	2218	4.4056	AluSx	SINE/Alu
DR2	AGGTGAGGAGGGCA	3657	1.3575		
IR1	AAGTCAGAGACCT	3822	1.0977		
DR4	GGATCACCTGAGGTCA	4426	3.1543	AluSx	SINE/Alu
DR4	ATGTGATAAAAGGTCA	5943	1.6589		
DR3	GGAACTCCTTGACCC	6262	1.3796	L1PA4	LINE/L1
ER1	AGAGCTGAGTTCA	8674	1.4353	L1PA4	LINE/L1
ER6	TGAACTAGTTTACAGTCA	10878	1.3791	L1PA4	LINE/L1
ER8	TGAACTGGATAAAGAGTGAA	11869	2.1048	L1P2	LINE/L1
DR4	AGACCTTAAATGACCT	12776	2.7327	L1P2	LINE/L1
DR1	TGGACTGTGAACT	14900	2.5632		
Human *CYP3A43*					
IR0	AGACCATGGCCT	1407	1.0179		
IR1	GAGCCACAGTCCT	3967	0.7625	MER65A	LTR/ERV1
DR3	AGGTCAGGGAGGGCC	4654	0.8757	MER65A	LTR/ERV1
IR1	AGGCCAGGGACCC	6380	1.1765	MER65A	LTR/ERV1
DR4	AAGTCAAAGGAGGTCA	8270	2.5709		
DR4	TGACCCCTACTACCCC	10056	0.8672		
DR2	TGACCTTGTGATCC	10735	2.1178	MER65A	LTR/ERV1
ER6	TGGACTTTTCAGAAGTCA	11019	1.8818		
DR2	AGGCCAGGAGTTCA	11316	1.6541	MER65A	LTR/ERV1
DR4	TGCCCTCCACTCAACT	11533	1.3123		
DR0	AGGTCAGGGTCA	12460	3.3143		

*refer to the position relative the transcriptional start site. [&]sites belong to the three known function modules (proximal ER6, XREM and CLEM) in human *CYP3A4*.

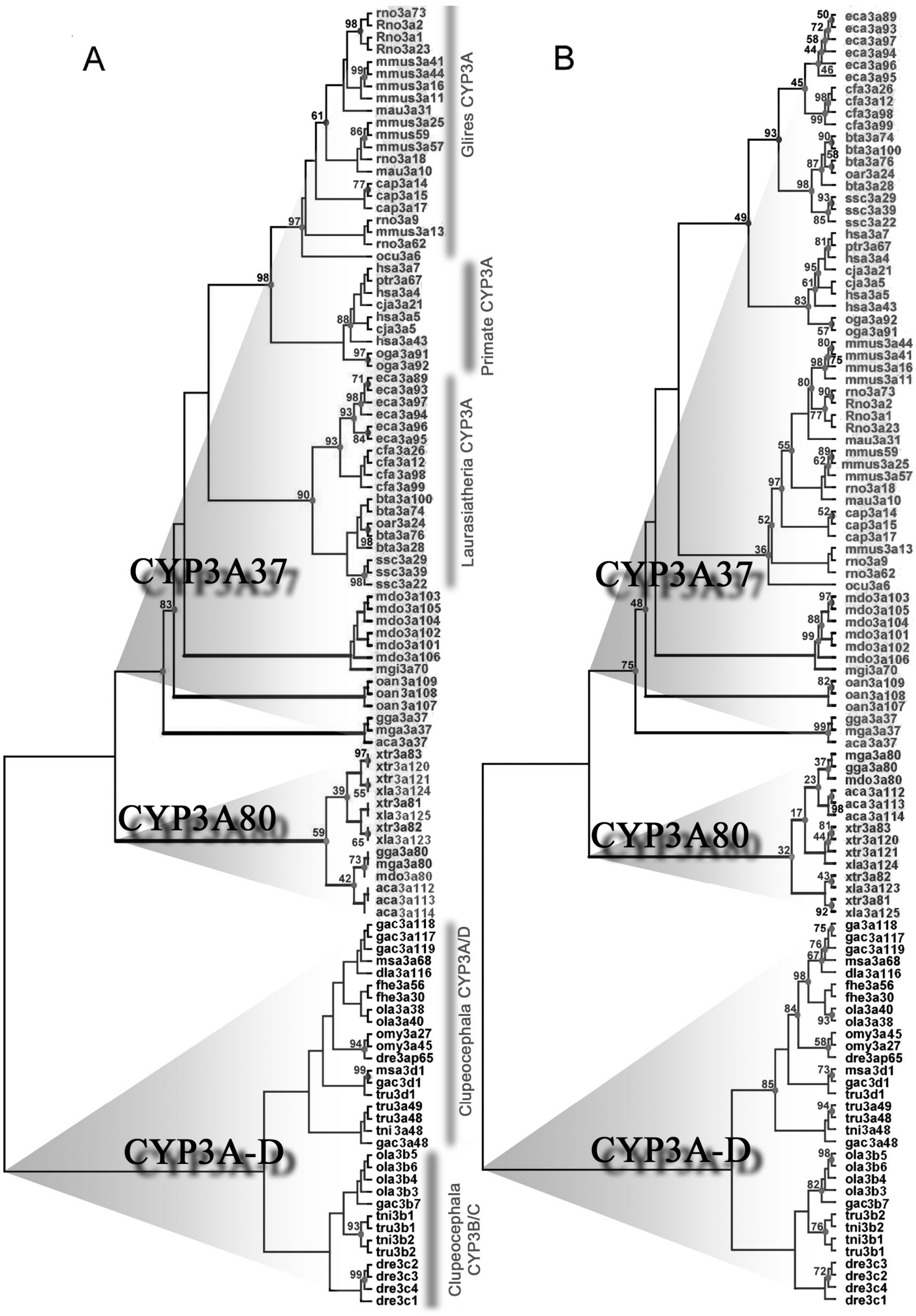


Fig. 8.1. Phylogeny of 107 CYP3A protein sequences reconstructed by Bayesian (A) and Maximum Likelihood (B) methods. Genes of different clades (*CYP3A37, CYP3A80* and *CYP3A-D*) are indicated. Only the posterior probabilities smaller than 100 are shown at each node labeled by dots. Boot strap values smaller than 1000 are shown as two-digital values at each node labeled by dots. The CYP3A clusters used for functional divergence detection (see section 2.4 and 3.5) were indicated. hsa (*Homo sapiens*), ptr (*Pan troglodytes*), cja (*Callithrix*

jacchus), oga (*Otolemur garnettii*), ami (*Alligator mississippiensis*), cfa (*Canis familiaris*), ssc (*Sus scrofa*), bta (*Bos taurus*), oar (*Ovis aries*), ocu (*Oryctolagus cuniculus*), mau (*Mesocricetus auratus*), mmus (*Mus musculus*), rno (*Rattus norvegicus*), cap (*Cavia porcellus*), msa (*Micropterus salmoides*), eca (*Equus caballus*) mdo (*Monodelphis domestica*), oan (*Ornithorhynchus anatinus*), gga (*Gallus gallus*), mga (*Meleagris gallopavo*), mgi (*Macropus giganteus*), aca (*Anolis carolinensis*), xtr (*Xenopus tropicalis*), dre (*Danio rerio*), tru (*Takifugu rubripes*), ola (*Oryzias latipes*), gac (*Gasterosteus aculeatus*), tni (*Tetraodon nigroviridis*), dla (*Dicentrarchus labrax*), omy (*Oncorhynchus mykiss*), fhe (*Fundulus heteroclitus*).

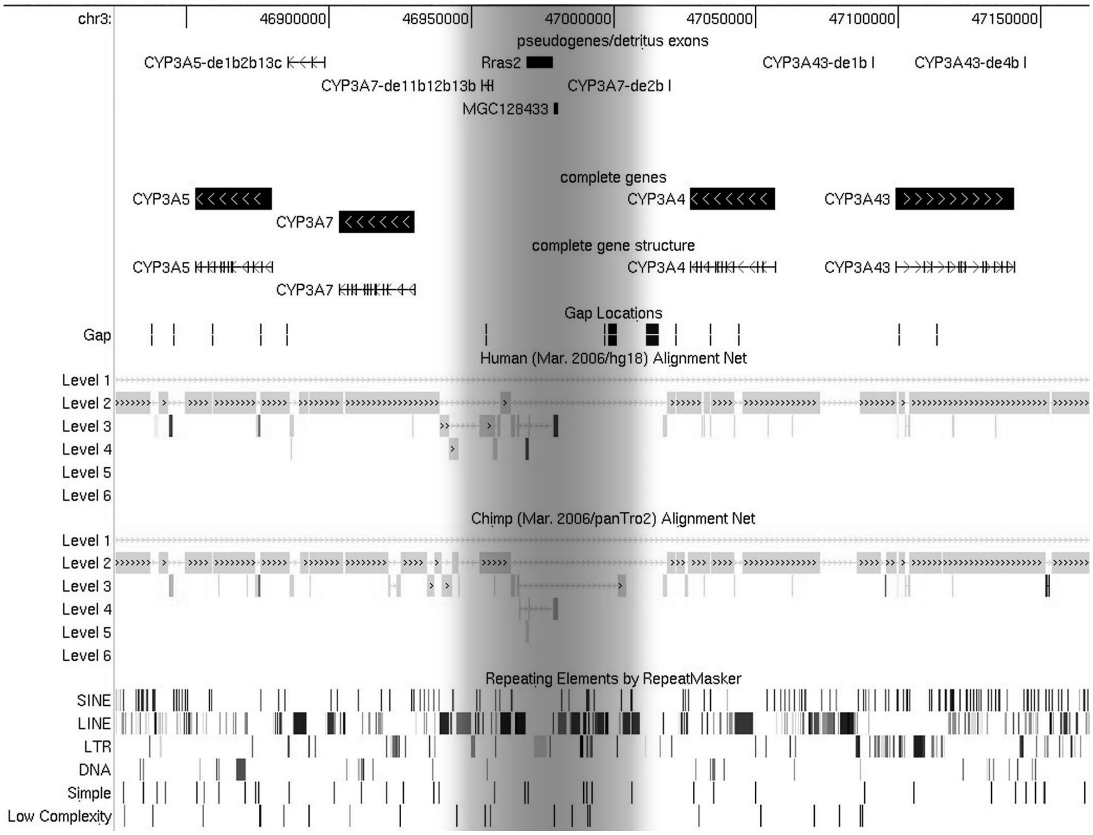


Fig. 8.2. Structure of the rhesus *CYP3A* locus. The position of the locus on rhesus chromosome 3 (rheMac2) are indicated by numbers at the top. Genes, pseudogenes, and detailed gene structure annotation are displayed in their corresponding tracks. Vertical and horizontal lines represent exons and introns, respectively. Arrows indicate orientation of genes. Gaps and repetitive elements are indicated by black boxes in Gap Location and Repeating Elements by RepeatMasker track. The Human and Chimp Alignment Net track shows the condensed best human/rhesus chain that indicates the regions that have orthologs in human and chimpanzee genome. The boxes represent ungapped alignments and the lines represent gaps. The insertion, compared with human locus, between *CYP3A7* and *CYP3A4* is shadowed.

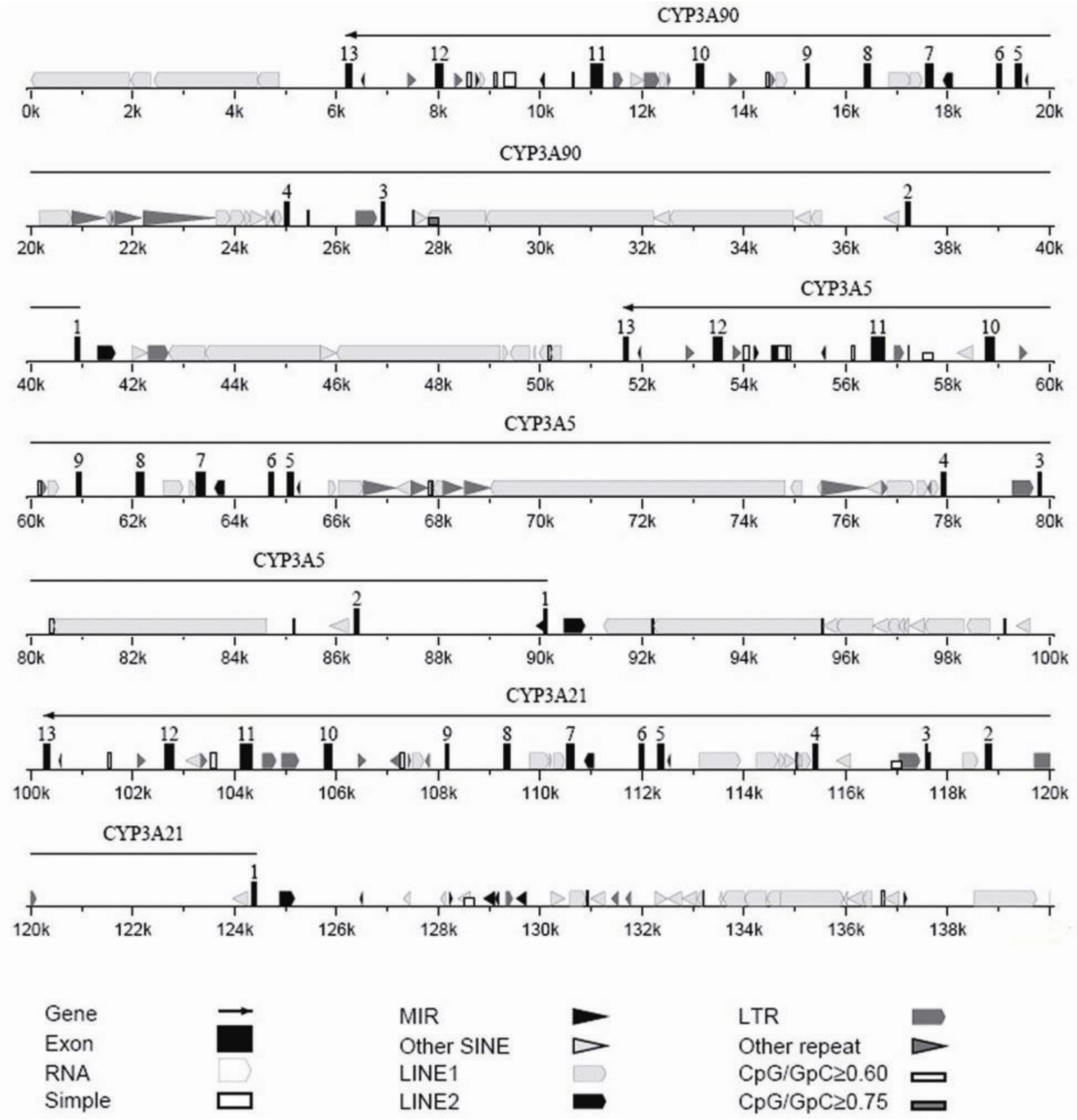

CYP3A90
13 12 11 10 9 8 7 6 5
0k 2k 4k 6k 8k 10k 12k 14k 16k 18k 20k
CYP3A90
4 3 2
20k 22k 24k 26k 28k 30k 32k 34k 36k 38k 40k
CYP3A5
1 13 12 11 10
40k 42k 44k 46k 48k 50k 52k 54k 56k 58k 60k
CYP3A5
9 8 7 6 5 4 3
60k 62k 64k 66k 68k 70k 72k 74k 76k 78k 80k
CYP3A5
2 1
80k 82k 84k 86k 88k 90k 92k 94k 96k 98k 100k
CYP3A21
13 12 11 10 9 8 7 6 5 4 3 2
100k 102k 104k 106k 108k 110k 112k 114k 116k 118k 120k
CYP3A21
1
120k 122k 124k 126k 128k 130k 132k 134k 136k 138k
Gene
Exon
RNA
Simple
MIR
Other SINE
LINE1
LINE2
LTR
Other repeat
CpG/GpC≥0.60
CpG/GpC≥0.75

Fig. 8.3. Structure of the marmoset *CYP3A* locus. The annotations of gene structure and repeat elements are displayed using PipMaker (Schwartz et al. 2000).

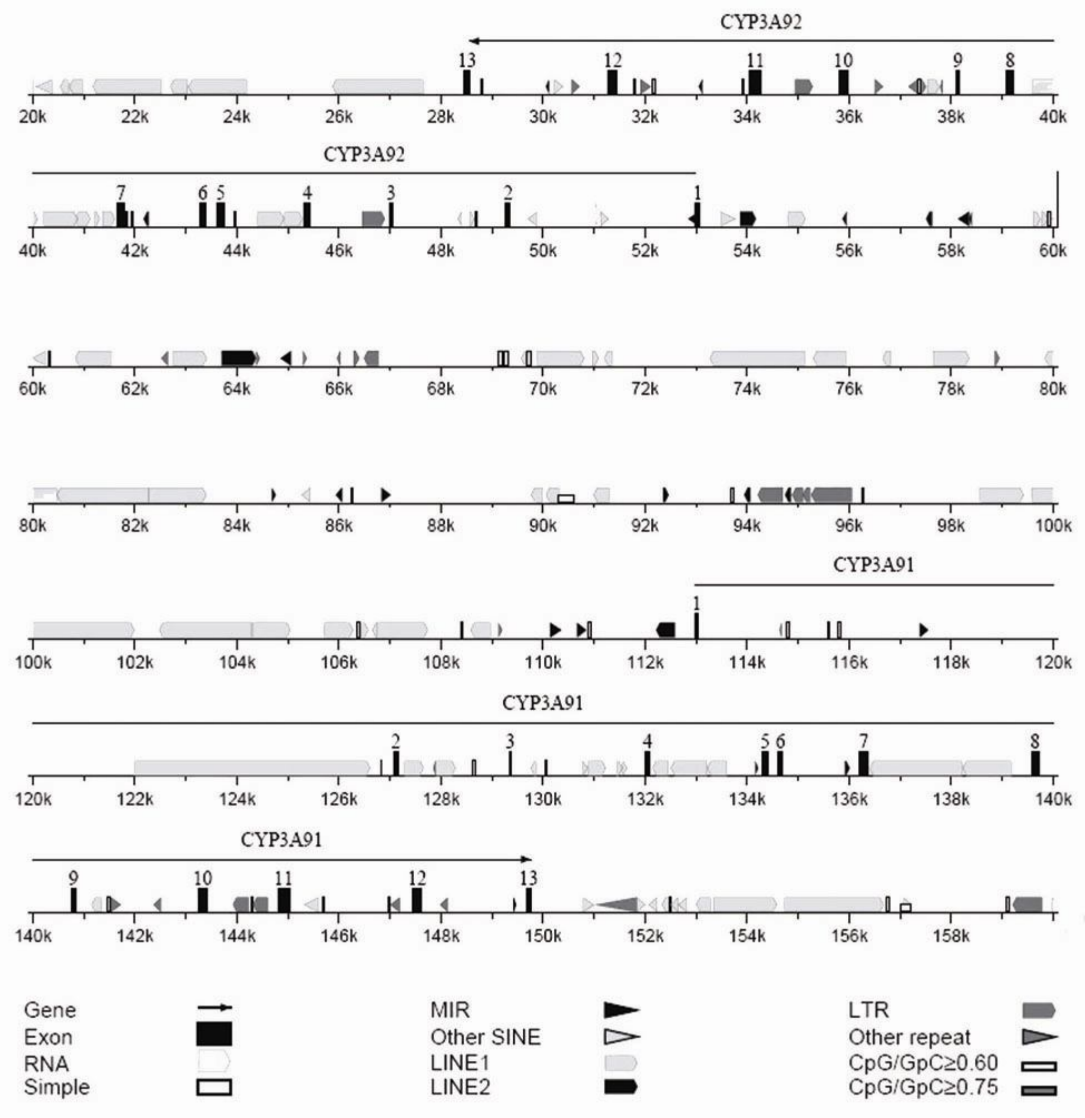


Fig. 8.4. Structure of a part of the galago *CYP3A* locus. The annotations of gene structure and repeat elements are displayed using PipMaker (Schwartz et al. 2000).

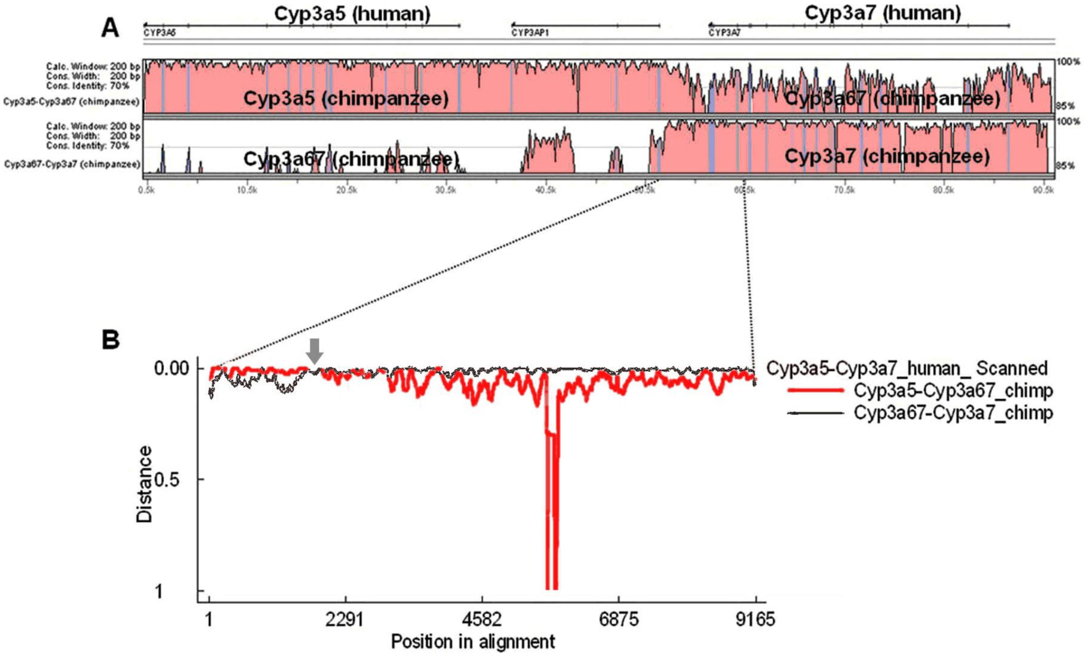
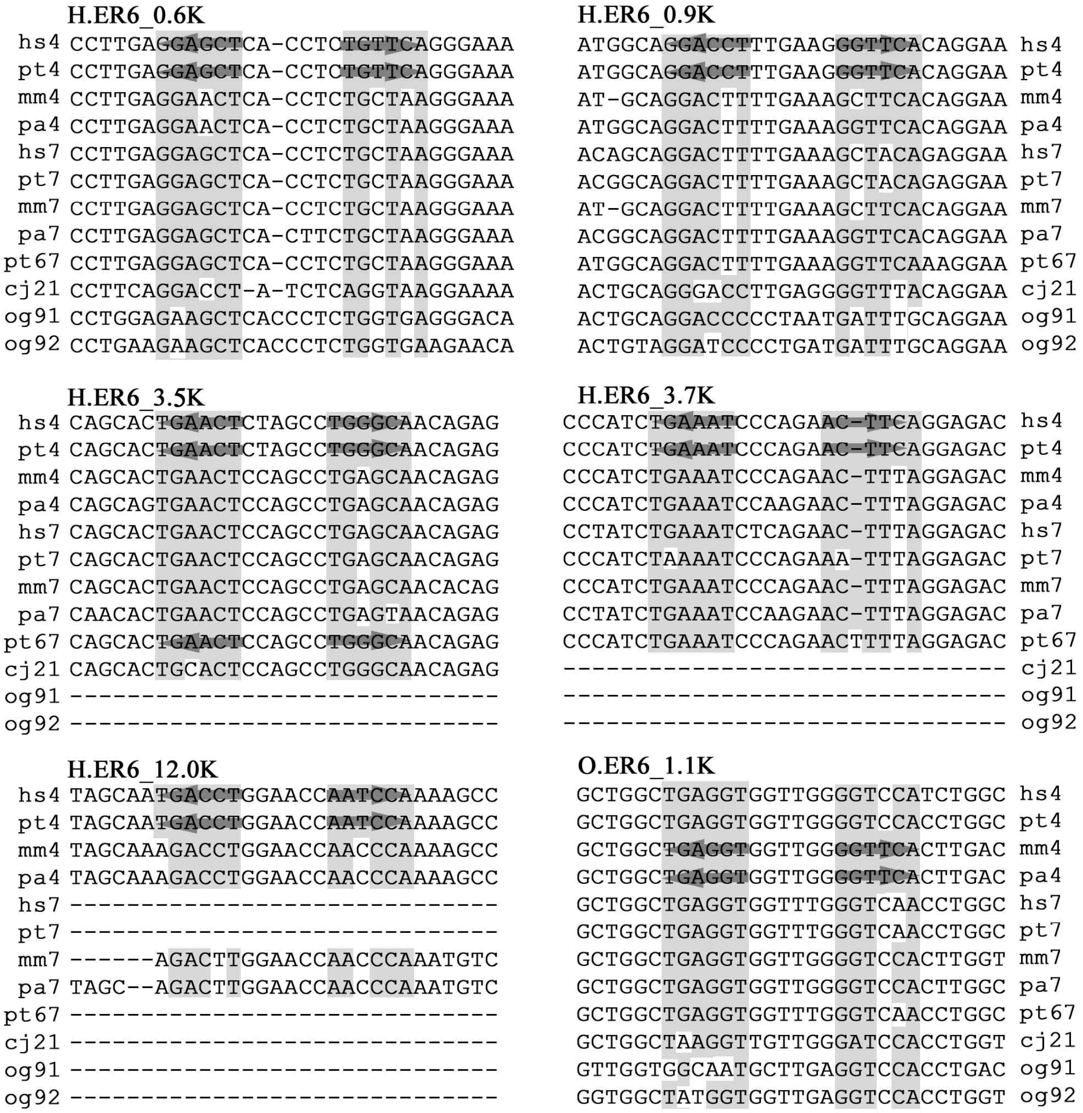


Fig. 8.5. *CYP3A67* loss in human lineage by homologous unequal crossover. Human *CYP3A5-CYP3A7* cassette was aligned with chimpanzee *CYP3A5-CYP3A67* and *CYP3A67-CYP3A7* cassette, respectively, using Multi-LAGAN. The 3' end of human *CYP3A7* and chimpanzee *CYP3A67/7* cassette were further analyzed using distance plot implemented in RDP (Martin and Rybicki 2000). (A) The similarity plots of human *CYP3A5-CYP3A7* cassette with chimpanzee *CYP3A5-CYP3A67* and *CYP3A67-CYP3A7* cassette. (B) The distance plots of 3' end human *CYP3A7* cassette with that of chimpanzee *CYP3A67* and *CYP3A7* cassette. The human *CYP3A5-CYP3A7* cassette appears to result from a homologous recombination event which linked *CYP3A5* directly with *CYP3A7* and lead to the loss of *CYP3A67*. The breakpoints were mapped to a region ~5 kb downstream of *CYP3A7* and *CYP3A67*.



```
     H.ER6_0.6K                                  H.ER6_0.9K
 hs4 CCTTGAGGAGCTCA-CCTCTGTTCAGGGAAA     ATGGCAGGACCTTTGAAGGGTTCACAGGAA hs4
 pt4 CCTTGAGGAGCTCA-CCTCTGTTCAGGGAAA     ATGGCAGGACCTTTGAAGGGTTCACAGGAA pt4
 mm4 CCTTGAGGAACTCA-CCTCTGCTAAGGGAAA     AT-GCAGGACTTTTGAAAGCTTCACAGGAA mm4
 pa4 CCTTGAGGAACTCA-CCTCTGCTAAGGGAAA     ATGGCAGGACTTTTGAAAGGTTCACAGGAA pa4
 hs7 CCTTGAGGAGCTCA-CCTCTGCTAAGGGAAA     ACAGCAGGACTTTTGAAAGCTACAGAGGAA hs7
 pt7 CCTTGAGGAGCTCA-CCTCTGCTAAGGGAAA     ACGGCAGGACTTTTGAAAGCTACAGAGGAA pt7
 mm7 CCTTGAGGAGCTCA-CCTCTGCTAAGGGAAA     AT-GCAGGACTTTTGAAAGCTTCACAGGAA mm7
 pa7 CCTTGAGGAGCTCA-CTTCTGCTAAGGGAAA     ACGGCAGGACTTTTGAAAGGTTCACAGGAA pa7
pt67 CCTTGAGGAGCTCA-CCTCTGCTAAGGGAAA     ATGGCAGGACTTTTGAAAGGTTCAAAGGAA pt67
cj21 CCTTCAGGACCT-A-TCTCAGGTAAGGAAAA     ACTGCAGGGACCTTGAGGGGTTTACAGGAA cj21
og91 CCTGGAGAAGCTCACCCTCTGGTGAGGGACA     ACTGCAGGACCCCCTAATGATTTGCAGGAA og91
og92 CCTGAAGAAGCTCACCCTCTGGTGAAGAACA     ACTGTAGGATCCCCTGATGATTTGCAGGAA og92

     H.ER6_3.5K                                  H.ER6_3.7K
 hs4 CAGCACTGAACTCTAGCCTGGGCAACAGAG    CCCATCTGAAATCCCAGAAC-TTCAGGAGAC hs4
 pt4 CAGCACTGAACTCTAGCCTGGGCAACAGAG    CCCATCTGAAATCCCAGAAC-TTCAGGAGAC pt4
 mm4 CAGCACTGAACTCCAGCCTGAGCAACAGAG    CCCATCTGAAATCCCAGAAC-TTTAGGAGAC mm4
 pa4 CAGCAGTGAACTCCAGCCTGAGCAACAGAG    CCCATCTGAAATCCAAGAAC-TTTAGGAGAC pa4
 hs7 CAGCACTGAACTCCAGCCTGAGCAACAGAG    CCTATCTGAAATCTCAGAAC-TTTAGGAGAC hs7
 pt7 CAGCACTGAACTCCAGCCTGAGCAACAGAG    CCCATCTAAAATCCCAGAAA-TTTAGGAGAC pt7
 mm7 CAGCACTGAACTCCAGCCTGAGCAACACAG    CCCATCTGAAATCCCAGAAC-TTTAGGAGAC mm7
 pa7 CAACACTGAACTCCAGCCTGAGTAACAGAG    CCTATCTGAAATCCAAGAAC-TTTAGGAGAC pa7
pt67 CAGCACTGAACTCCAGCCTGGGCAACAGAG    CCCATCTGAAATCCCAGAACTTTTAGGAGAC pt67
cj21 CAGCACTGCACTCCAGCCTGGGCAACAGAG    ------------------------------- cj21
og91 ------------------------------    ------------------------------- og91
og92 ------------------------------    ------------------------------- og92

     H.ER6_12.0K                                 O.ER6_1.1K
 hs4 TAGCAATGACCTGGAACCAATCCAAAAGCC     GCTGGCTGAGGTGGTTGGGGTCCATCTGGC hs4
 pt4 TAGCAATGACCTGGAACCAATCCAAAAGCC     GCTGGCTGAGGTGGTTGGGGTCCACCTGGC pt4
 mm4 TAGCAAAGACCTGGAACCAACCCAAAAGCC     GCTGGCTGAGGTGGTTGGGGTTCACTTGAC mm4
 pa4 TAGCAAAGACCTGGAACCAACCCAAAAGCC     GCTGGCTGAGGTGGTTGGGGTTCACTTGAC pa4
 hs7 ------------------------------     GCTGGCTGAGGTGGTTTGGGTCAACCTGGC hs7
 pt7 ------------------------------     GCTGGCTGAGGTGGTTTGGGTCAACCTGGC pt7
 mm7 ------AGACTTGGAACCAACCCAAATGTC     GCTGGCTGAGGTGGTTGGGGTCCACTTGGT mm7
 pa7 TAGC--AGACTTGGAACCAACCCAAATGTC     GCTGGCTGAGGTGGTTGGGGTCCACTTGGC pa7
pt67 ------------------------------     GCTGGCTGAGGTGGTTTGGGTCAACCTGGC pt67
cj21 ------------------------------     GCTGGCTAAGGTTGTTGGGATCCACCTGGT cj21
og91 ------------------------------     GTTGGTGGCAATGCTTGAGGTCCACCTGAC og91
og92 ------------------------------     GGTGGCTATGGTGGTTGAGGTCCACCTGGT og92
```

Fig. 8.6. Sequence alignments of the ER6 elements specific to hominoid (human/chimpanzee) or OWM (rhesus and baboon). Conserved nucleotides in half sites are grey shadowed. Predicted ER6 elements are indicated by arrows representing each of the half sites. Gene names are shortened (e.g., hs4: Homo sapiens *CYP3A4*, pt7: Pan troglodyte *CYP3A7*). hs (*Homo sapiens*), pt (*Pan troglodytes*), mm (*Macaca mulatta*), pa (*Papio anubis*), cj (*Callithrix jacchus*), og (*Otolemur garnettii*).