CYP3 phylogenomics: evidence for positive selection of CYP3A4 and CYP3A7

Huan Qiua, Stefan Taudienb, Holger Herlynb, Jürgen Schmitzd, Yuan Zhoue, Guopei Chenf, Roberta Robertof, Mariano Rocchif, Matthias Plätzeerc and Leszek Wojnowskia

Objective CYP3A metabolizes 50% of currently prescribed drugs and is frequently involved in clinically relevant drug interactions. The understanding of roles and regulations of the individual CYP3A genes in pharmacology and physiology is incomplete.

Methods Using genomic sequences from 16 species we investigated the evolution of CYP3 genomic loci over a period of 450 million years.

Results CYP3A genes in amniota evolved from two ancestral CYP3A genes. Upon the emergence of eutherian mammals, one of them was lost, whereas, the other acquired a novel genomic environment owing 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: (i) on CYP3A7 early in hominoid evolution, which was accompanied by a restriction of its hepatic expression to fetal period and (ii) 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.

Conclusions CYP3A7 and CYP3A4 may have acquired catalytic functions especially important for the evolution of hominoids and humans, respectively. Pharmacogenetics and Genomics 18:53–66 © 2008 Wolters Kluwer Health | Lippincott Williams & Wilkins.

Keywords: comparative genomics, CYP3A4, CYP3A7, phylogenomics, positive selection

Introduction CYP3A4, CYP3A5, CYP3A7, and CYP3A43 form a gene cluster on human chromosome 7. Protein expression in vivo has been demonstrated for CYP3A4, CYP3A5, and CYP3A7. The highest levels of CYP3A proteins are found in the small intestine and in the liver, in agreement with their well-known role in the detoxification of exogenous compounds. The physiological importance of the expression in several other organs, such as prostate and kidney (CYP3A5) or adrenal gland (CYP3A7) is less well understood, except it may involve hydroxylation of steroids and of other physiological compounds. CYP3A4 is the most abundant hepatic CYP3A isoform in adults. CYP3A5 is expressed only in adult carriers of CYP3A5*1 alleles (~70% Africans, ~30% Asians, and ~10% Central Europeans), in most cases at levels lower than those of CYP3A4. CYP3A7 is the only CYP3A expressed in fetal livers, but it is also found in adult carriers of CYP3A7*1C alleles (~15% Central Europeans). CYP3A43 is considered a pseudogene, based on a low level of mostly aberrant transcripts, although bacteria-expressed protein exhibits some activity [1].

CYP3A metabolizes, at least in part, every second drug currently in use [2]. The reported substrate spectra for CYP3A5 and CYP3A7 are smaller in comparison with CYP3A4, although this judgment is affected to an unknown extent by the variability in the critically important CYP3A expression and reconstitution conditions [1]. CYP3A proteins are frequently involved in clinically relevant drug interactions [2], owing to the combination of their broad substrate spectrum with one or more of the following three factors: (i) pharmacokinetics of CYP3A drug substrates is affected by the largely inherited [3] but unpredictable individual CYP3A4 expression level in the liver and in the small...
The variability may be further enhanced by (ii) CYP3A induction and (iii) inhibition by certain drugs and dietary constituents. The transcription-mediated CYP3A induction, for example, by prescription-free drugs such as St John's wort, may increase the metabolic clearance and attenuate the therapeutic efficacy of a CYP3A substrate [4]. The inducing effects of drugs are mediated by the xenobiotics sensors pregnane X receptor and, to a lesser extent, constitutive androstane receptor. The modulation of CYP3A expression by endogenous compounds is exemplified by the effect of vitamin D, which is mediated by vitamin D receptor [5]. Inhibition, for example, by grapefruit juice has an opposite effect and may increase the concentration of a CYP3A substrate to a toxic level [2].

For all these reasons, 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 [5]. 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. The evolution of a gene's sequence among human populations, or among species, reflects in part adaptations to changing physiology or environment. We [6] and others [7] recently reported an unusually strong positive selection signal in the CYP3A locus in non-African human populations. CYP3A10 loci in these ethnic groups contain low numbers of long-range haplotypes, consistent with their rapid expansion owing to positive selection. Salt–water homeostasis [7] and rickets [6] have been proposed as the underlying selection factors, consistent with the expression of CYP3A5 in the kidney and with the activity of CYP3A4 toward vitamin D, respectively. The identification of selecting factors and genes is notoriously difficult, owing to the confounding impact of population bottlenecks, founder effects, or periods of rapid population expansion [8], but it is medically relevant. Indeed, recent episodes of positive selection frequently represent a response to potentially harmful changes in diet or environment. Some of these factors may be at work today, consistent with the association of CYP3A alleles with several current human diseases, notably cancers [9].

Together with CYP3B, 3C, and 3D, CYP3A subfamily belongs to the CYP3 family, believed to have diverged from other CYP genes at least 800 million years ago [10]. CYP families are defined by at least 40% amino acid sequence similarity, whereas the corresponding number for subfamilies is 55% [11]. Previous phylogenetic analyses of CYP3 have been mainly performed on cDNA or protein sequences [12,13]. This had certain unavoidable limitations, as, in addition to true relatedness, similarities among cDNA sequences are strongly affected by species-specific selection and by genomic recombination events. Moreover, owing to limited sampling, CYP3 evolution in basally diverging amniota and fish-like vertebrates was largely unknown.

Whole genome sequencing of a rapidly increasing number of species makes possible addressing the evolution of gene families by comparative genomic approaches. Using genomic sequences from 16 vertebrate species, we reconstruct in the present work the genomic evolution of CYP3A and of the related subfamilies CYP3B–D during the last ~450 million years. This is complemented by the investigation of gene conversion events and 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. On the basis of this work, we investigated the evolutionary regime that has shaped the contemporary primate CYP3A protein-coding sequence, using the ratio of nonsynonymous (amino acid altering) and synonymous (silent) substitution rates \((d_{n}/d_{s} = K_{a}/K_{s} = \omega)\) as a measure.

Methods and materials
Genomic and cDNA sequences
Genomic assemblies of 16 species were accessed through either University of California Santa Cruz (UCSC) [14] or ENSEMBL (Ensembl is a joint project between EMBL-EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes) [15] genome browsers or National Center for Biotechnology Information (NCBI) Genbank database. The investigated species were human (Homo sapiens, hg18), chimpanzee (Pan troglodytes, panTro2), rhesus (Macaca mulatta, rhMac2), 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 CYP3A locus sequences from olive baboon (Papio anubis, Genbank no. AC141417). CYP3 sequences available from NCBI Genbank and D. Nelson's Cytochrome P450 Homepage (http://drnelson.utmem.edu/CytochromeP450.html) were retrieved and mapped in each vertebrate genome by basic local alignment search tool (BLAST).
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 [16]. 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 bacterial artificial chromosome (BAC) clones were isolated by screening the CHORI-259 and CHORI-256 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, Foster City, California, USA).

cDNA cloning of primate CYP3A genes
Lever samples of chimpanzee, orangutan, and rhesus were obtained from Max–Planck Institute for Evolutionary Anthropology, Leipzig, Germany; and liver samples of olive baboon and hamadryas baboon (Papio anubis) were obtained from Deutsches Primatenzentrum Göttingen. The RNA was extracted using TRIzol following a standard protocol. Total RNA from the liver of additional rhesus individual was purchased from BioChain Institute, Inc. (Hayward, California, USA). Reverse transcription was performed using SuperScript transcriptase (Invitrogen, Karlsruhe, Germany). Marmoset cDNA was obtained from University of Göttingen, Germany. Primers and PCR conditions of primate CYP3A cloning are described in the Supplemental Table S1. The products of reverse transcriptase PCRs (RT-PCRs) were sequenced, either directly or following AT cloning, by GENterprise (Mainz, Germany).

Phylogenetic analysis
One hundred and seven CYP3 amino acid sequences from 31 species were aligned (available upon request) using ClustalX [17] and were 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 [18]. Bayesian inference was performed using program MrBayes [19,20], which estimates posterior probabilities of clade support using Metropolis-coupled Monte Carlo-Markov Chain method (MC²). 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 10 million generations. For each run, three chains were heated and one was cooled with a temperate parameter of 0.20. Rate variation across sites was approximated using a four-category γ distribution and the proportion of invariable sites was estimated from the data assuming a uniform prior distribution. Trees were sampled every 100 generations and following a burn-in of 1000 generations. A 95% majority rule consensus tree was generated to calculate posterior probability values. Maximum likelihood analysis was done using phylogenetic inferences using maximum likelihood (PHYML) [21,22]. The initial tree was determined by neighbor-joining (BIONJ). Rate variation across sites was approximated using a four-category γ distribution. The tree topology, branch lengths, γ-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 CYP3A genes (Supplemental Fig. S3) was inferred from full-length coding cDNA sequences. Bayesian inference was performed with HKY + G substitution model following the procedure described above for protein sequences analysis. To verify the primate CYP3A phylogeny, rare genomic changes (RGCs) were investigated in the 20 CYP3A genes (Supplemental Fig. S4) with complete genomic sequences. The sequences were aligned by using Multi-limited area global alignment of nucleotide (LAGAN) [23] and shared orthologous intronic retroposed elements and random indels (minimum two nucleotides) were screened as described [24]. The presence/absence of these elements was then mapped on the topology of a sequence-based phylogenetic tree (Supplemental Fig. S4).

Gene conversion
GENECOV program (http://www.math.wustl.edu/~sawyer/ geneconv/) [25] 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 speed 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 [23] aligned CYP3A genomic sequences. 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.

**Likelihood ratio tests for positive selection**

Evolution of primate CYP3A sequences was assessed using the ratio of nonsynonymous (amino acid altering) and synonymous (silent) substitution rates \( (\omega = d_{\text{a}}/d_{\text{s}} = \text{Ka/Ks} = \omega) \) as a measure. As a very conservative estimate, positive (Darwinian) selection \( (\omega > 1) \) is assumed, if the nonsynonymous substitution rate exceeds the synonymous substitution rate, owing 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, owing to their detrimental effect on fitness. Finally, neutral evolution \( (\omega = 1) \) indicates that nonsynonymous substitutions do not affect individual fitness. The analysis was performed using the maximum likelihood approach implemented in phylogenetic analysis by maximum likelihood (PAML) package version 3.15 [26–28]. 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 tree used was taken from above analyses. Site-specific 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 3 x 4 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 lineagespecificity of \( \omega \) (LRT III). For each of the LRTs, twice the log-likelihood difference between alternative and null model was compared with critical values from the log-likelihood difference between alternative and null model as a measure. As a very conservative estimate, positive (Darwinian) selection \( (\omega > 1) \) is assumed, if the nonsynonymous substitution rate exceeds the synonymous substitution rate, owing 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, owing to their detrimental effect on fitness. Finally, neutral evolution \( (\omega = 1) \) indicates that nonsynonymous substitutions do not affect individual fitness. The analysis was performed using the maximum likelihood approach implemented in phylogenetic analysis by maximum likelihood (PAML) package version 3.15 [26–28]. 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 tree used was taken from above analyses. Site-specific 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 3 x 4 matrix. We performed two parallel likelihood ratio tests (LRTs) for the null models, that is, nearly neutral model M1a and beta null model M7, respectively. To avoid local optima, M8 was run twice with initial \( \omega \) values smaller or larger than 1. Candidate sites for positive selection were pinpointed using the Bayes empirical Bayes approach. Only sites with a minimum support of 0.95 under M2a or M8 were further considered. LRT III compared free-ratio model and one-ratio model. For details of the implementation of each model, see Refs. [26–28]. To pinpoint candidate sites of positive selection along branches of particular interest (Ho7 + Hs4, see Fig. 5), we inferred ancestral sequences using Baseml (model = HK85, cleandata = 0). Subsequently, we identified nonsynonymous exchanges by pairwise comparison of the sequences at the ends of Ho7 and Hs4, respectively.

**Results**

**Phylogenomics of CYP3 loci**

A total of 25 loci (Fig. 1) containing full-length CYP3 genes (Supplemental Table S2) were identified in the 16 genomes queried by BLAST with representative CYP3 sequences from each subclade. CYP3B, C, and D genes were found only in fish species. In most vertebrates, apparently intact CYP3 genes are located within one of two CYP3 homologous regions, CYPHR1 and CYPHR2. CYP3HR1 (SDKI-CYP3-FOXK1-KIA0415-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-ATPS22-ZNF3-YHAP-AKAT-A1-ME4-TRIME4-GJIE1-AZIP1). 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 fish species contains CYP3C or CYP3D genes, in stickleback together with CYP3A genes. Several additional CYP3-containing genomic regions were

<table>
<thead>
<tr>
<th>LRTs</th>
<th>Model</th>
<th>j</th>
<th>Test statistics</th>
<th>np</th>
<th>Estimates of parameters</th>
<th>Sites with highest support from BEB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT I</td>
<td>M1a (nearly neutral)</td>
<td>-5643.60</td>
<td>2Δl= 11.44, cv= 9.21</td>
<td>2</td>
<td>P_0 = 0.95; P_1 = 0.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(site-specific)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M2a (positive selection)</td>
<td>-5637.88</td>
<td>2Δl= 11.44, cv= 9.21</td>
<td>2</td>
<td>P_0 = 0.95; P_1 = 0.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT II</td>
<td>M7 (beta)</td>
<td>-5645.72</td>
<td>2Δl= 15.78, cv= 13.82</td>
<td>2</td>
<td>P = 0.15, q = 0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(site-specific)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M8 (beta&amp;\alpha)</td>
<td>-5637.83</td>
<td>2Δl= 15.78, cv= 13.82</td>
<td>4</td>
<td>P = 0.50, q = 0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT III</td>
<td>M0 (one ratio)</td>
<td>-5692.18</td>
<td>2Δl= 53.06, cv= 46.19</td>
<td>1</td>
<td>P = 0.09, P_\alpha = 0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(branch-specific)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Free ratio</td>
<td>-5665.65</td>
<td>2Δl= 53.06, cv= 46.19</td>
<td>33</td>
<td>P = 0.09, P_\alpha = 0.11</td>
<td></td>
</tr>
</tbody>
</table>

BEB, Bayes empirical Bayes approach; cv, critical value from \( \chi^2 \) distribution given the difference in the number of free parameters of the models compared; j, 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 lineagespecificity of \( \omega \); np, number of free parameters; 2Δl, twice the log-likelihood difference of the models compared; P, level of significance.
found outside CYPHR1 and CYPHR2 in the frog and in the fish species (Fig. 1). Thus, the CYP3B subfamily is contained in a syntenic region (KCNH-CYP3B) in all fish species but zebrafish, whereas all frog CYP3A genes were found in two loci specific for amphibians.

**Gene conversion**

We identified a total of nine potential gene pairs that might have been affected by gene conversion involving the protein-coding regions. These included CYP3 from rat, dog, pig, medaka, and some primate species (Supplemental Table S3). 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 hominid ancestor, as it is detectable in human, chimpanzee, and orangutan sequences. This is also supported by a 15-bp deletion found in CYP3A4 and CYP3A7 genes in these three species, and absent from these genes in rhesus and baboon, and from CYP3A67 and CYP3A5 genes in all species (Fig. 2). Moreover, analysis of full-length genomic sequences revealed a number of additional gene conversion events restricted to intron sequences (data not shown).
Phylogeny of ray-finned fish and amniota CYP3

A schematic representation of a phylogenetic tree based on 107 CYP3 protein-coding sequences is given in Fig. 3. The entire tree is provided as a Supplemental Fig. S1. Phylogenetic analysis using Bayesian and maximum likelihood methods produces similar trees, the only difference being the relationships of rodents, primates, and Laurasiatheria CYP3A within the well-supported eutherian CYP3A clade. We analyzed in more detail fish and non-eutherian amniota, as underrepresented in previous analyses [12,13]. The highest CYP3 diversity has been achieved in fish species. Examination of the expressed sequence tag (EST) data from several other fish species (Supplemental Table S4) indicates that CYP3B and CYP3C must have existed at least as early as Percomorpha and Cypriniformes ancestors, respectively. Although the posterior probability and bootstrap support is low (Supplemental Fig. S1), all extant amnioid CYP3A genes form two distinct groups (Fig. 3 and Supplemental Fig. S1). The first group (‘CYP3A80 clade’ in Fig. 3) includes bird CYP3A80, three anolis CYP3A genes, all frog CYP3A, and one CYP3A gene from opossum. The second group (‘CYP3A37 clade’ in Fig. 3) comprises bird CYP3A37, one anolis CYP3A, most CYP3A genes from opossum and platypus, and all eutherian CYP3A genes. These data together indicate 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 in Eutheria are derived from an ancient CYP3A37 ortholog.

Genomics and phylogenetics of primate CYP3A

Four and five CYP3A genes have been reported, respectively, within the human [29] and chimpanzee CYP3A loci [13]. 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, contains four complete genes (CYP3A5, 7, 4, and 43), and several pseudogenes composed of detritus exons [11] in an organization resembling that found in the human locus (Fig. 4, see Supplemental Fig. S2 for details). An ~40-kb genomic insertion consisting mainly of low copy number repeats was found between rhesus CYP3A4 and CYP3A7. Sequence analysis of a baboon CYP3A-containing BAC (Genbank no. AC141417) revealed a similar insertion between baboon CYP3A7 and CYP3A4 genes (data not shown). The marmoset CYP3A locus (Genbank accession nos. EF600757

Pharmacogenetics and Genomics 2008, Vol 18 No 1

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
and EF600758) contains three genes (CYP3A5, 90, and 21) in tandem (Fig. 4). 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 (data not shown). Galago CYP3A locus (Genbank accession nos. EF600755 and EF600756) contains at least three tandem CYP3A genes of which one is in the opposite orientation to the others (Fig. 4).

We also examined orangutan CYP3A-related WGS data in 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 reverse transcriptase PCR (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. We note 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 [13,30].

The reconstructed process of primate CYP3A locus evolution shown in Fig. 4 is in agreement with primate
CYP3A phylogeny (Supplemental Figs. S3 and S4). Galago CYP3A resides on the basal position of the primate phylogenetic tree. Marmoset CYP3A21 shares a common ancestor with catarrhine CYP3A4, whereas marmoset CYP3A5 and CYP3A90 share a common ancestry with catarrhine CYP3A5. Human CYP3A, CYP3A4, and CYP3A7 orthologs are restricted to Hominidae. No human CYP3A genes have one-to-one 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 primate CYP3A genes (Supplemental Fig. S4). Only two conflicting RGCs were found, of which one is covered by the gene conversion shown in Fig. 2.

**Evidence for positive selection**

According to LRT I and II, the log likelihood value of positive selection model M2a and beta<sup>o</sup> model M8 was significantly higher than the log likelihood value, inferred under assumption, of nearly neutral model M1a and betanull model M7, respectively (Table 1). Consequently, positive selection can be assumed for single sites of
the 18 catarrhine CYP3A genes studied. According to parameter estimates from M2a and M8, about one tenth of sites are under moderate positive selection ($\omega = 2.3–2.6$). Estimates from posterior probability provided significant support of positive selection for codons 437, 478, and 479 ($P_o > 0.95$), when taking M8 (Table 1). Though not significant, support from posterior probability was still high for the same sites from M2a ($P_o > 0.87–0.92$). Likewise, the free-ratio model fits the data significantly better than the one-ratio model M0, thus suggesting lineage-specificity of $\omega$ (LRT III in Table 1). Out of the 33 branches of the analyzed phylogeny, six branches showed evidence of positive selection (Fig. 5), with the highest $\omega$ values for branch Ho7 (CYP3A7 in the hominoid stem line; $\omega = infinite$) and Hs4 (human CYP3A4; $\omega = infinite$). 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. 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 show nonsynonymous substitutions along Ho7 (H28R, S29T, V50A, F57Y, F74I, G77C, R78Q, V81M, E333K, A337T, N437S, R484L). The corresponding number for Hs4 was 6 (R54H, R78Q, I129L, I224T, R478S, T489V). The concurrent absence of synonymous exchanges renders these codon site 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 1). Moreover, most branches are characterized by $\omega$ values less than 1 and, hence, negative selection (Fig. 5).

Selection within catarrhine CYP3A phylogeny. The branches with $\omega$ 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), and Ponpy (Pongo pygmaeus).
Discussion
Our combination of comparative genomics, phylogenetics, and selection analysis appears to be a valuable extension of techniques usually applied to investigate enzyme function, such as radiograph analysis, spectroscopic measurements, site-directed mutagenesis, mechanism-based inhibition, and photoaffinity labeling. The last ~450 million years of CYP3 evolution displays several remarkable features. Most CYP3 genes of the same clade in the phylogenetic tree build 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 [31]. The oppositely (head-to-head or tail-to-tail) oriented genes, if any, are always the most distant related, 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 the common role of inverted duplication in the early expansion of CYP3 loci.

The highest diversification of CYP3 has taken place in fish species, which are the only ones containing CYP3B through 3D subfamilies. This is suggestive of a function of CYP3B–3D genes in the adaptation to specific aquatic environments, consistent with the expression of CYP3C1 in the gill and skin of zebrafish [32]. CYP3 diversification in the fish species was in a few cases independent from their syntenic relationships at genomic level. Although CYP3HR1 typically harbors CYP3A genes, in zebrafish it contains four CYP3C genes with organ expression different from CYP3A31 [32]. Conversely, the zebrafish CYP3A65 is located apart from the CYP3HR1. A similar situation is observed in fugu and stickleback CYP3 genes, in which both CYP3D genes are found in CYP3HR1.

CYP3A genes from amniota form two distinct groups in the phylogenetic tree (‘CYP3A37’ and ‘CYP3A80’ in Fig. 3). 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. Although opossum CYP3A101–106 and platypus CYP3A107–109 result from independent duplications of an ancestral CYP3A37 homolog, the opossum CYP3A80 gene arose from an ancestral CYP3A80 homolog.

Upon the emergence of eutherian mammals, the ancestral CYP3A80 ortholog got lost, whereas, 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 for some of CYP3A genes, eutherian CYP3HR1 remains remarkably intact, indicating no structural instability. It is tempting to speculate that the novel genomic location, that is, CYP3HR2, may have been selected following a rare translocation event from CYP3HR1 owing to certain fitness advantages. It is known that some genomic locations affect gene expression [33] and protein evolution [34,35]. CYP3A is expressed in placenta, where it may play a fetus-protecting role by metabolizing steroids and foreign substrates [36].

Evolution of primate CYP3A
Primate CYP3A loci appear to have evolved very different in Strepsirrhini and New World monkey compared with Old World monkeys and Hominoida. The head-to-head orientation of the galago CYP3A (CYP3A91 and CYP3A92) resembles that of CYP3A43 and CYP3A44 in Catarrhini (Fig. 4). The sister group relationship of Strepsirrhini (i.e. galago) CYP3A and all anthropoid CYP3A in the reconstructed phylogeny, however, 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 (Supplemental Figs S3 and S4). Contrary to the repeated duplication of CYP3A21-like genes (CYP3A44, 7, 67) in catarrhines, platyrrhines (New World monkey) expanded only CYP3A45, as evidenced by CYP3A locus structure of the marmoset. Evidence of a marmoset CYP3A43 ortholog was found neither by screening of a genomic BAC library, nor by the extension of the locus assembly based on the sixfold coverage WGS data until the flanking gene TRIM4 (data not shown). As CYP3A43 is located on the basal position of anthropoid CYP3A phylogeny (Supplemental Figs. S3 and S4), it must have been created prior to the split of catarrhines and platyrrhines, and therefore is likely to be lost in this species. 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 [12,13]. Although CYP3A21 and CYP3A4 exhibit similarities in transcriptional regulation [37], the 5’ upstream regions of marmoset CYP3A45 and CYP3A90 notably differ from that of human CYP3A5, owing to independent retrotransposable element insertions (data not shown). 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.

CYP3A67 arose through a duplication of CYP3A7, early in the evolution of Hominidae, as 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 European and 38 African (Bantu) DNA samples.
sequence evolution along branches Ho7 and Hs4. In addition, we detected two episodes of accelerated conservation of their original biochemical functions. In primate CYP3A protein evolution, consistent with the dominant role of purifying (i.e. negative) selection in the CYP3A4/CYP3A7 group, first changing into 108L within the CYP3A7 lineage and reverting to 108F 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 human CYP3A4 and CYP3A7, yet a strong selection signal is detected only in branch Ho7. This apparent contradiction can be explained by the pleisomorph F residue, common for the CYP3A21/CYP3A4/CYP3A7/CYP3A467 group, first changing into 108L within the CYP3A7 lineage and reverting to 108F owing to the selection on Ho7. Together with 215F and five 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 interacts with the electron donors cytochrome b5 or P450 reductase [42].

The identity of the amino acid at position 108 in CYP3A4 affects the activity toward midazolam [43], aflatoxin B1 [44], the steroids testosterone and progesterone [44,45], and lapachole [46]. The importance of the 215F, 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 [47], but its mutagenesis affected neither the activity nor the cooperativity of the enzyme [48,49].

Of the 18 amino acids 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 six 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.

Positive selection and its functional implications

Our 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, we detected two episodes of accelerated sequence evolution along branches Ho7 and Hs4. In general, variation of $\omega$ among lineages is not considered to be sufficient evidence for adaptive evolution, as it can as well result from relaxation of selective constraint [41]. In case of Ho7 and Hs4 evidence for positive selection, however, is rather strong: first, the $\omega$ estimate for Ho7 and Hs4 is infinite, that is as high as it can be. Second, the results from branch-specific analysis (free-ratio model) are supported by pairwise sequence comparison that identified 18 and 6 candidate sites of positive selection along Ho7 and Hs4. Taken together, it appears justified to us to assume positive selection for Ho7 and Hs4, instead of relaxation of 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 Hyllobatidae (gibbon). The exact evolutionary branch will be identified only after the cloning of gibbon

(data not shown). Owing to the large gaps within the CYP3 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 CYP3A61 [30]. In contrast, our data suggest an unequal crossover with the breakpoints located downstream of CYP3A67 and CYP3A67 (Fig. 4 and Supplemental Fig. S5).

Similar to human CYP3A43 [29,38], 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 resulting in a frame shift. We found the same deletion in one of the three rhesus samples investigated (data not shown). These observations are consistent with an ongoing pseudogenization of CYP3A43, although a protein-independent function [39] cannot be excluded. Owing to the multitude of possible mechanisms, gene loss or pseudogenization are more likely outcomes of a mutation than gain-of-function gene variants [40]. The loss of CYP3A43 (humans) and CYP3A43 (marmoset), and the pseudogenization of CYP3A43 (human, rhesus, and chimpanzee), may therefore represent adaptive responses. On the other hand, the coding region of CYP3A43 appears to have evolved under strong purifying constraint, as evidenced by the low $\omega$ values for the terminal branches leading to these genes [40].

The loss or pseudogenization are more likely outcomes of a collagen mutation than gain-of-function gene variants [40]. The loss or pseudogenization are more likely outcomes of a collagen mutation than gain-of-function gene variants [40].

Of the 18 amino acids identified as 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 six 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 human CYP3A4 and CYP3A7, yet a strong selection signal is detected only in branch Ho7. This apparent contradiction can be explained by the pleisomorph F residue, common for the CYP3A21/CYP3A4/CYP3A7/CYP3A467 group, first changing into 108L within the CYP3A7 lineage and reverting to 108F owing to the selection on Ho7. Together with 215F and five 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 interacts with the electron donors cytochrome b5 or P450 reductase [42].

The identity of the amino acid at position 108 in CYP3A4 affects the activity toward midazolam [43], aflatoxin B1 [44], the steroids testosterone and progesterone [44,45], and lapachole [46]. The importance of the 215F, 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 [47], but its mutagenesis affected neither the activity nor the cooperativity of the enzyme [48,49].

Of the 18 amino acids selected in hominid CYP3A7 (Ho7), eight 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 [45]. 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 [50]. The
exact molecular determinants of the reduced interaction with b5 are unknown. The hydrophobic region of the N terminus may also be important for entering of some substrates into the active site [51].

Codons 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 the activity by 80–90% and changed the regioselectivity toward aflatoxin B1 metabolites [44]. The substitution of CYP3A4 L479 with T479 (like in CYP3A5) had a similar, reducing effect on the production of the same aflatoxin B1 metabolites. Additionally, it halved the hydroxylase activity of the mutant protein toward testosterone. The mutation of the CYP3A4 479L to 479F (like in CYP3A7) changed the product profile of 7-hexoxyccoumarin [52]. 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 [53]. 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 are 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, as both selection events were lineage-specific rather than geographic region-specific. Human CYP3A7 is the only CYP3A that shows a recent positive selection of its protein sequence. The verification of this hypothesis will require identification of functionally relevant differences in CYP3A4 regulatory elements between humans and chimpanzees.

Most interestingly, our results allow hypothesizing about the reason for the large interindividual 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, as both processes can improve fitness. For example, an increased activity of an enzyme such as CYP3A4 could be brought about through structural changes or by increased expression. Increasing evidence for a coupling between the evolution of protein sequences and their expression levels is found [61,62]. 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 [63]. Taken together, the interindividual variability in CYP3A4 expression in humans may reflect the accelerated evolution of human CYP3A4 protein and its 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. The verification of this hypothesis will require identification of functionally relevant differences in CYP3A4 regulatory elements between humans and chimpanzees.

Acknowledgements
The authors thank Dr Wolfgang Enard (Max–Planck Institute for Evolutionary Anthropology in Leipzig, Germany) for the chimpanzee, orangutan, and rhesus liver samples; Dr Christian Roos (German Primate Center, Göttingen) for olive baboon and hamadryas baboon liver samples; Professor Victor Armstrong (University Göttingen) for common marmoset cDNA; Daniela Werler and Oliver Müller for skillful technical assistance; Markus Schilhabel for bioinformatics support; Professor Hans Zischler (University Mainz) for many useful suggestions throughout the project; and Professor David Nelson (University of Tennessee) for help and comments on sequence data. This work was supported in part by the Deutsche Forschungsgemeinschaft Grant WO505/2-2. The authors acknowledge the support of M.P. and S.T. and L.W. by the Nationales Genomforschungsnetz (NGFN grants 01GR0504 and 01GS0421, respectively).

Supplement data
Supplementary data are available at the Pharmacogenetics and Genomics journal online (www.pharmacogeneticsandgenomics.com).

References


