

Rapid Birth–Death Evolution Specific to Xenobiotic Cytochrome P450 Genes in Vertebrates

James H. Thomas*

Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

Genes vary greatly in their long-term phylogenetic stability and there exists no general explanation for these differences. The cytochrome P450 (*CYP450*) gene superfamily is well suited to investigating this problem because it is large and well studied, and it includes both stable and unstable genes. *CYP450* genes encode oxidase enzymes that function in metabolism of endogenous small molecules and in detoxification of xenobiotic compounds. Both types of enzymes have been intensively studied. My analysis of ten nearly complete vertebrate genomes indicates that each genome contains 50–80 *CYP450* genes, which are about evenly divided between phylogenetically stable and unstable genes. The stable genes are characterized by few or no gene duplications or losses in species ranging from bony fish to mammals, whereas unstable genes are characterized by frequent gene duplications and losses (birth–death evolution) even among closely related species. All of the *CYP450* genes that encode enzymes with known endogenous substrates are phylogenetically stable. In contrast, most of the unstable genes encode enzymes that function as xenobiotic detoxifiers. Nearly all unstable *CYP450* genes in the mouse and human genomes reside in a few dense gene clusters, forming unstable gene islands that arose by recurrent local gene duplication. Evidence for positive selection in amino acid sequence is restricted to these unstable *CYP450* genes, and sites of selection are associated with substrate-binding regions in the protein structure. These results can be explained by a general model in which phylogenetically stable genes have core functions in development and physiology, whereas unstable genes have accessory functions associated with unstable environmental interactions such as toxin and pathogen exposure. Unstable gene islands in vertebrates share some functional properties with bacterial genomic islands, though they arise by local gene duplication rather than horizontal gene transfer.

Citation: Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3(5): e67. doi:10.1371/journal.pgen.0030067

Introduction

Genes in animal and plant gene families vary greatly in their phylogenetic stability. Some genes persist as a single copy over a wide phylogenetic range of species, with few or no gene duplications or losses on different evolving lineages. Other genes undergo frequent duplication and loss in a process called birth-death evolution [1,2]. Gene duplication and subsequent divergence of the duplicate copies underlie the formation of gene families and are thought to be an important source of genetic complexity and evolutionary change [3–5]. Though many specific examples of stable and unstable genes have been documented, there exists no general explanation for their different patterns of evolution. To explore the evolutionary basis for these patterns, I sought a gene family with characteristics that make it amenable to detailed analysis and interpretation. The vertebrate cytochrome P450 (*CYP450*) gene superfamily is well suited for the following reasons. First, it is a large family with both phylogenetically stable members and unstable members. Second, the biochemical and organismal functions of most genes in the family have been extensively studied in humans and rodents. Third, their protein products are large and have a well-conserved domain content, providing abundant information for studies of molecular evolution. Finally, strong interest in the family has resulted in relatively high quality annotations of gene structure in a wide range of vertebrate species.

The human genome contains approximately 60 *CYP450* genes, which encode membrane-bound oxidase enzymes that act on a wide variety of substrates (recent reviews include [6–8]). The *CYP450* enzymes can be divided into those that act on endogenous small molecules and those that act on xenobiotic compounds. The endogenous-substrate enzymes function in biosynthesis or catabolism of steroids, sterols, retinoids, prostaglandins, and fatty acids. The xenobiotic-substrate enzymes are expressed primarily in the liver and epithelial tissues and function to defend against environmental toxins and carcinogens. These xenobiotic *CYP450* genes have been studied extensively in humans because they comprise one of the major (and most polymorphic) activities that determine drug half-life and are thus important in pharmaceutical development [6,7]. Xenobiotic *CYP450* enzymes have also been studied extensively because in some

Editor: Barbara J. Trask, Fred Hutchinson Cancer Research Center, United States of America

Received: December 6, 2006; **Accepted:** March 14, 2007; **Published:** May 11, 2007

A previous version of this article appeared as an Early Online Release on March 14, 2007 (doi:10.1371/journal.pgen.0030067.eor).

Copyright: © 2007 James H. Thomas. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: *CYP450*, cytochrome P450; TBLASTN, translated BLAST nucleotide

* To whom correspondence should be addressed. E-mail: jht@u.washington.edu

Author Summary

Genes vary greatly in their long-term phylogenetic stability, and there exists no general explanation for these differences. Stable genes persist as a single copy over a wide range of distantly related species, whereas unstable genes undergo frequent duplication and loss in a process called birth-death evolution. The vertebrate cytochrome P450 (*CYP450*) gene superfamily includes many genes that are present in a single copy in species ranging from teleost fish to mammals and other groups of genes that undergo active birth-death evolution across the same species. The author found that nearly all stable *CYP450* genes encode enzymes known to function in the synthesis and degradation of steroid and retinoid hormones (and related molecules). These hormones function in core developmental pathways in vertebrates. In contrast, most unstable *CYP450* genes encode enzymes that detoxify foreign small molecules (called xenobiotics—foreign biochemicals). In addition, many of the unstable *CYP450* genes are subject to natural selection to change their amino acid sequence over time (positive selection), probably in response to changes in xenobiotic exposure. These findings suggest that stable and unstable genes differ in their rates of birth-death evolution, because stable genes have core endogenous functions whereas unstable genes respond to changing environmental conditions.

cases they convert exogenous compounds into active carcinogens, presumably as a negative side-effect of broad substrate specificity [9]. Approximately 22 of the human *CYP450* enzymes are thought to act primarily on endogenous substrates and approximately 15 other *CYP450* enzymes are thought to act primarily on xenobiotic substrates (Figure S1). A few of the remaining enzymes are reported to have activity toward both endogenous and xenobiotic substrates; it is unclear whether one or the other activity (or both) is their primary function. Finally, there are about ten *CYP450* enzymes with no known activities, most of which have been identified only recently from human genome sequence.

To study the molecular evolution of the *CYP450* superfamily, I considered for analysis about 20 vertebrate genomes in various stages of draft sequencing, assembly, and annotation. Of these, sequence coverage, assembly, and annotation of 11 genomes were sufficiently complete for my needs. Of these 11, *Pan troglodytes* was discarded because of its extreme similarity to human, leaving ten nearly complete genomes: six placental mammals, one bird, one amphibian, and two teleost fish. Analysis of gene stability among these genomes reveals that the rate of birth-death evolution among genes acting on xenobiotic substrates is much higher than among those acting on endogenous substrates.

Results

Phylogenetic Trees

As starting material, I gathered complete sets of predicted *CYP450* genes from the genomes of human (*Homo sapiens*), rhesus macaque (*Macaca mulatta*), dog (*Canis familiaris*), cow (*Bos taurus*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), clawed frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), and pufferfish (*Takifugu rubripes*). All known and predicted splice forms were screened to identify a single canonical coding sequence from each gene for molecular evolutionary analysis (see Materials and Methods). Figure S1 shows a maximum-likelihood protein tree derived from all

628 of these sequences, marked to indicate known or probable functions based on the human proteins. A representative segment of the tree is shown in Figure 1. In the tree, *CYP450* enzymes that act on endogenous substrates are generally represented by a single gene in each organism, and the phylogeny of that gene approximates the known species phylogeny. In some cases there are two related proteins in one or both fish species, presumably resulting from the known whole-genome duplication on the fish lineage [10,11]. In all but one case, each orthologous group of genes is strongly separated on the tree from all other *CYP450* genes, indicating that they became functionally specialized well before the divergence of teleost fish from mammals (the single exception is described below). In sharp contrast, nearly all *CYP450* enzymes that act on xenobiotic substrates are encoded by genes that have undergone active duplication and loss throughout vertebrate evolution, reflected in multiple species- and lineage-specific gene expansions. The tree segment in Figure 1 includes two endogenous-substrate enzymes and one class of xenobiotic-substrate enzymes; they are typical of the complete tree (Figure S1). Table 1 summarizes gene numbers and variance across species for three groups of *CYP450* genes.

Among the endogenous-substrate ortholog groups, an expected protein from a specific species was occasionally absent, usually from cow, chicken, or frog. One example is found in Figure 1 (a *CYP46A1* cholesterol 24-hydroxylase ortholog is missing from cow). Most of these apparent gene losses probably result from incomplete sequence assemblies or gene annotations. Such apparent gene losses were found most often in the cow, chicken, and frog genomes presumably as a result of relatively immature assemblies and annotations of these genomes. In several cases, translated BLAST nucleotide (TBLASTN) searches of the cognate genome clearly indicated the likely presence of part or all of an unpredicted gene ortholog (see Materials and Methods; Tables S1 and S2). There were also a few cases of apparent species-specific gene duplications outside of fish. For example, there were two predicted rat genes encoding *CYP51A1*-like proteins (Figure S1). Although these cases may include undiagnosed pseudogenes, it is reasonable to suppose that occasional gene duplications may be tolerated in these groups. Finally, there was one case in which enzymes that act on endogenous substrates (corticosteroid precursors) appear to have become specialized recently and specifically in terrestrial vertebrates (*CYP11B1* steroid 11- β -hydroxylase and *CYP11B2* aldosterone synthase), probably in connection with resistance to dehydration [12]. Nevertheless, the general picture among endogenous-substrate *CYP450* genes is one of striking phylogenetic stability, with very low rates of gene duplication and loss across jawed vertebrates.

Among *CYP450* enzymes with xenobiotic substrates the picture was dramatically different. Almost without exception, one or more species or lineages had multiple genes that grouped in the tree with one or no genes from other species. The sample shown in Figure 1 is typical of the complete tree. These differences in gene content are far too extensive to be accounted for by incomplete genome sequence and annotation. In particular, the human and mouse genomes are especially mature in sequence and annotation, and the pattern of gene duplication and loss is robustly apparent in comparing these two genomes (e.g., Figure 1; Table 2).

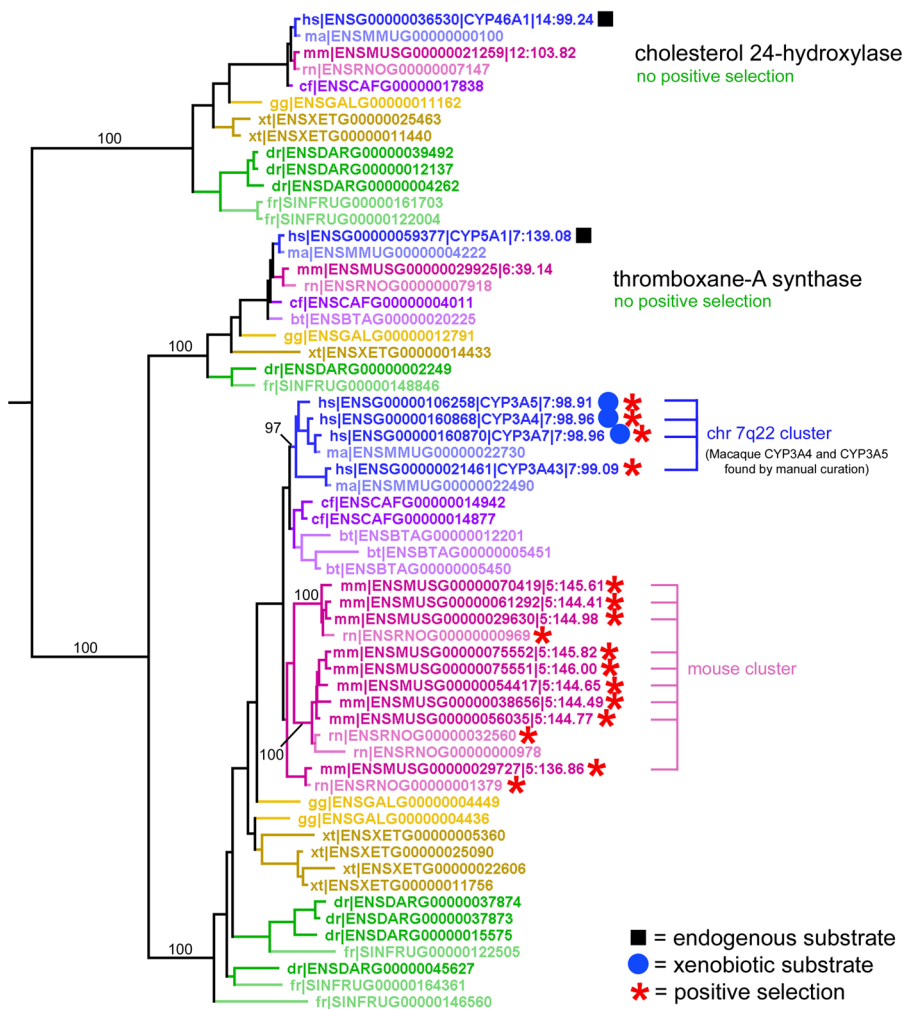


Figure 1. Excerpt from Tree of Vertebrate CYP450 Proteins

Shown is one segment of 61 proteins from a maximum-likelihood protein tree of all 628 CYP450 proteins from 10 vertebrate species. The complete tree is shown in Figure S1. Protein names are color coded by species and include a genus-species prefix (e.g., “hs” is *H. sapiens*). Human and mouse protein names end with the chromosome and position (in Mb) of the corresponding gene. Solid black squares mark human proteins with a known endogenous substrate. Solid blue circles mark human proteins with known xenobiotic substrates. Red asterisks mark genes for which positive selection was detected; tests for positive selection were performed only on primate and rodent genes. Other markings indicate enzyme activity, gene clusters in human (blue), and mouse (pink), and specific groups tested where positive selection was not found (green). Numbers on tree branches show percent bootstrap support for all branches important for interpretation. Almost half of the missing cow *CYP46A1* ortholog is present in the Btau3.1 assembly based on a TBLASTN search (Chromosome 21, plus strand at ~58.4 Mb) (see Table S1). Though not apparent from this tree segment, clear expansions in the number of *CYP450* genes are present on other parts of the complete tree for fish, frog, and chicken (Figure S1). doi:10.1371/journal.pgen.0030067.g001

Extensive bootstrap testing on several subtrees supported the occurrence of independent gene duplications and deletions on both the mouse and human lineages (Figure S2).

In principle, genetic exchange among closely related genes within a species (concerted evolution) could explain some aspects of these patterns by causing lineage-specific clades of genes from sequence homogenization rather than recent duplication. In support of this possibility, a recent paper describes evidence that concerted evolution played a role in the divergence of the *CYP1A1* and *CYP1A2* genes in mammals from their bird orthologs [13]. However there are several reasons to think that this case is unusual. First, all of the syntenic *CYP450* gene clusters between human and mouse contained different numbers of genes, indicating that gene duplication or loss must occur (two examples are shown in Figures S3 and S4). Second, except for a few of the most closely-related gene pairs, nucleotide divergence among

duplicate genes was substantial, nucleotide changes were distributed widely across the coding sequence, and regions of higher nucleotide similarity usually coincided with regions of stronger purifying selection in the family in general (unpublished data). Finally, in many cases the tree relationships were inconsistent with concerted evolution being a dominant influence on sequence relatedness (for example, Figure S5). These results suggest that concerted evolution plays, at most, a minor role in shaping evolution in the family as a whole.

Unstable Genes Are Clustered in the Human and Mouse Genomes

Most human and mouse genes for CYP450 enzymes are scattered widely across their genomes, but there are some exceptions. Clusters of genes are found at human chromosome bands 1p33, 7q22, 10p13, 10q26, and 19q13. In each of

Table 1. Gene Group Counts

Gene Group	Human	Macaque	Dog	Cow	Mouse	Rat	Standard Deviation/Mean
<i>CYP2</i>	17	18	11	19	43	33	0.51
<i>CYP3/4</i>	16	13	5	14	24	15	0.42
Stable (18 sets)	18	15	17	18	18	17	0.07
Total CYP450	59	55	40	60	92	74	NA

CYP450 genes were divided into major groups for analysis: the unstable *CYP2* family, the unstable *CYP3*, and *CYP4* families (which are closely related), and 18 stable families. The *CYP1* family was not included because it is small and contains both stable and unstable genes. Each data entry is the number of potentially functional genes in each class. The last column is the standard deviation of gene number among the species divided by the mean. Gene counts, especially in macaque, dog, cow, and rat, are likely to be revised up with continued genome assembly and annotation. Gene counts may include a few recent pseudogenes when they encode a nearly complete *CYP450* protein sequence. NA, not available.
doi:10.1371/journal.pgen.0030067.t001

these five human clusters, the *CYP450* genes are adjacent, with no other confirmed genes interspersed among them. Within each cluster, all the genes encode closely related proteins with many cases of apparent gene duplication since the split from the rodent lineage. Strikingly, nearly all of the *CYP450* enzymes with known xenobiotic substrates were found in these gene clusters (e.g., see Figure 1 chromosome positions). The remaining genes in these clusters were also phylogenetically unstable but are not yet implicated in xenobiotic detoxification. Each of the human clusters had a syntenic counterpart cluster in mouse (Figures S1, S3, and S4). Very few of the genes in these clusters could be assigned as one-to-one orthologs because of repeated duplication and deletion events on both lineages. These syntenic gene clusters must have originated from a shared ancestral gene or genes, with repeated local gene duplications and losses giving rise to lineage-specific groups of related genes. The patterns of divergence are consistent with single gene duplications that occurred at various times, though multigene duplications followed by loss of most of the duplicate genes cannot be ruled out. Genomic clustering of related genes was also apparent in the other eight genomes, but assembly quality was more problematic, and I did not investigate the patterns in detail. These results are consistent with increased rates of birth-death evolution specific to xenobiotic-substrate genes, with local gene duplications giving rise to clusters of related genes.

Many Unstable Genes Are Subject to Positive Selection

If exposure to xenobiotic compounds changed over evolutionary time, the *CYP450* detoxifier genes might be subject to positive selection driven by changing substrates. To test this possibility, I used maximum-likelihood codon analysis on paralogous groups of genes from many parts of the overall *CYP450* tree [14,15]. Most attention was given to tree regions where there was evidence of frequent recent gene duplication, both because this might indicate selection to diversify enzyme function and because this provided abundant sequence data for analysis. Specifically, 203 genes in clades with five or more closely-related genes were tested, including six groups from primates, nine from rodents, and six from a mixture of primates and rodents. Details of the method are given in Materials and Methods, and summary statistics for the results are given in Table S3. A total of five of the 21 tested sets showed highly significant evidence of positive selection. The specific amino acid sites under positive selection were analyzed for a set of 12 rodent *Cyp2d* genes, which showed the strongest signal of positive selection. In this

set, a Bayes-Empirical-Bayes method identified nine sites with strong evidence of positive selection ($p > 0.9$ for each site). The structural positions of these nine sites were compared by alignment with two closely related *CYP2* proteins for which crystal structures have been determined [16,17]. Both in the primary sequence alignment and in the 3-D protein structure, the nine sites under positive selection were clearly associated with substrate-binding or substrate-channel regions (Figures S6 and S7). In the protein structure, the distance from positive selection sites to a bound warfarin ligand was significantly smaller than for random sets of sites ($p \approx 0.0002$, see Materials and Methods). In addition to these nine sites, all three small indels among the 12 rodent genes coincided with regions of substrate binding (Figure S6). No method exists to analyze such indels for positive selection, but it is plausible that they affect substrate interactions and might be subject to such selection. These results are consistent with changing xenobiotic substrates driving the observed positive selection.

For comparison with the unstable genes, nine sets of genes from stable *CYP450* groups were also tested for positive selection (marked on Figure S1). In each case, the orthologous genes from the six mammals were tested as a group. Because there are almost no gene duplications in these groups, it was not possible to include paralogs in these sets, but the degree of divergence of the genes (total tree length) was in the same range as for the paralog groups, indicating that the maximum-likelihood analysis should have similar power [18]. No evidence of positive selection was found for any of the nine sets, as expected based on their function as enzymes acting on endogenous substrates.

Discussion

Phylogenetic instability among genes in families has been observed repeatedly, including among immunoglobulin genes [19], major histocompatibility complex genes [20], T-cell receptor genes [20,21], mammalian olfactory receptor genes [22], nematode chemosensory genes [15,23], globin genes [24,25], cichlid opsin genes [26,27], zinc-finger genes [28], F-box and MATH-BTB genes [29], and others. Selective pressure from environmental change or pathogens has been suggested as a driver of duplication and diversification for many of these examples, but alternatives are difficult to rule out. The vertebrate *CYP450* gene superfamily affords special insight into gene instability because probable biochemical functions of most of the genes are known and because it includes both stable and unstable genes. I found a strong

Table 2. Recent Probable Duplications

Family	Human Gene Number	Human Lineage Duplications	Mouse Gene Number	Mouse Lineage Duplications
CYP1	3	0	3	0
CYP2	17	5	43	24
CYP3	4	3	9	8
CYP4	12	6	15	7
Endogenous enzymes	22	1 ^a	22	1 ^a

CYP450 genes were divided into the four largest recognized sequence families (CYP1–4) and a diverse group that encodes enzymes that act on known endogenous substrates. Endogenous enzymes are marked on Figure S1 with filled black boxes. They are: CYP1B1, CYP4V2, CYP5A1, CYP7A1, CYP7B1, CYP8A1, CYP8B1, CYP11A1, CYP11B1, CYP11B2, CYP17A1, CYP19A1, CYP21A2, CYP24A1, CYP26A1, CYP26B1, CYP26C1, CYP27A1, CYP27B1, CYP39A1, CYP46A1, and CYP51A1.

^aThese apparent duplications may be an artifact (see tree).

doi:10.1371/journal.pgen.0030067.t002

correspondence between genes implicated in xenobiotic detoxification and evolutionary instability in the form of gene duplication and deletion. These findings strongly support the involvement of changing external selective pressure in high rates of birth-death gene evolution. Positive selection for amino acid change has been also been reported in many of the phylogenetically unstable gene families, suggesting that diversification is achieved by a combination of gene duplication and selection-driven divergence in sequence.

For other unstable gene families, lack of stable family members for comparison or lack of extensive functional information prevent the highly informed analysis possible for the *CYP450* genes. Nevertheless, a similar explanation for birth-death evolution is plausible for most or all such families. Immunoglobulin genes, major histocompatibility complex class I and class II genes, and T cell receptor genes function in pathogen defense, and an extensive literature documents gene duplications and deletions in these families during vertebrate evolution. Mammalian olfactory receptor genes and nematode chemosensory genes function in response to environmental chemical stimuli and might be subject to changes in desired ligand specificity or sensitivity. Lake Malawi cichlid fish have undergone duplications of opsin genes that have diverged to different spectral sensitivities, and expression of specific sets of these genes results in diverse adaptive visual properties, possibly driven by mate choice and foraging preferences [26,27]. In other cases the role of external selective forces in driving birth-death evolution is less obvious but plausible; for example the F-box and MATH-BTB genes in nematodes and plants have been hypothesized to function in pathogen defense, though the evidence is indirect [29].

Much less is known about specific functions of *CYP450* genes in nematodes and plants. Based on maximum-likelihood protein trees with three *Caenorhabditis* species (*briggsae*, *elegans*, and *remanei*) and with the plants *Arabidopsis thaliana* and *Oryza sativa* (rice), it is clear that birth-death evolution affects most members of the *CYP450* family in these groups as well. For example, only 7% of *Arabidopsis CYP450* genes and 38% of *C. elegans* genes had strict one-to-one orthologs within their groups (unpublished data). As in the human and mouse genomes, closely related unstable genes tended to be in genomic clusters. The exact degrees of stability and instability cannot be directly compared because the phylogenetic distances differ, and genome annotation qualities may vary.

A few *CYP450* genes in humans lack clear functional

information. The relative evolutionary stabilities of these genes may be predictive of general aspects of their functions. Among stable genes of unknown function, human *CYP20A1* and orthologs are well separated from all other *CYP450* genes on the protein tree and are encoded by a single gene in each of the ten species analyzed (top left of Figure S1). This pattern suggests an ancient and critical function that is shared among all the species. A less clear-cut but similar pattern is seen for an unnamed human *CYP27* (Chromosome 2) and the *CYP2U1* and *CYP2R1* genes. I speculate that these four enzymes will prove to be involved in biosynthesis or catabolism of endogenous substrates of importance throughout vertebrates. Consistent with this interpretation, *CYP20A1* expression is enriched in immune system cell types, *CYP2U1* is expressed broadly, and *CYP2R1* expression is enriched in testis and immune system [30]. Unstable genes for which I could find no described function include *CYP4A43*, *CYP4Z1*, *CYP4Z2* (possible pseudogene), *CYP4X1*, *CYP4A11*, *CYP4A22*, *CYP4F2*, *CYP4F3*, *CYP4F8*, *CYP4F11*, and *CYP4F12*. Among these, in addition to birth-death evolution, the *CYP4F* group is subject to significant positive selection, consistent with substrate-driven selection for amino acid change. I speculate that each of these genes functions in xenobiotic detoxification or some other process that acts on environmental substrates. Consistent with this interpretation, expression of most of these genes is enriched in liver (*CYP4A11*, *CYP4A22*, *CYP4F2*, *CYP4F3*, *CYP4F11*, and *CYP4F12*) or in tracheal tissue (*CYP4X1*) [30], which are tissues where other xenobiotic *CYP450* genes are known to function [6,7]. Finally, the evolution of the *CYP2W1* and *CYP2J2* genes is less clear, with duplications in some species but only a single gene in humans. *CYP2W1* has a broad expression pattern [30], and evolution of the group appears most consistent with a relatively stable function in mammals but extensive duplication and diversification in amphibians and teleost fish. *CYP2J2* expression is enriched in liver [30], and the group includes multiple genes in most species but only a single gene in humans and macaques.

Groups of closely related unstable *CYP450* genes are strongly clustered in the human and mouse genomes. A similar pattern of genome clustering is found in a number of other human and mouse gene families [31–33]. These patterns indicate that DNA duplications that persist in the genome occur predominantly locally. In contrast, about half of very recent large DNA duplications (segmental duplications) on the human lineage are interchromosomal, and some of the remainder occur nonlocally on the same chromosome [34–

37]. It is possible that these segmental duplications represent a gene duplication mechanism that has dramatically and recently increased on the primate lineage [34]. Alternatively, nonlocal segmental duplications may be more unstable than local duplications and fail to persist over longer evolutionary periods, so that they rarely give rise to persistent functional duplicate genes.

The combination of evolutionary instability and gene clustering has parallels with “genomic islands” found in bacteria. Genomic islands consist of blocks of DNA that are present in some bacterial isolates but absent from other closely related isolates (reviewed in [38]). Though the full picture is still developing, it is likely that most genes in genomic islands have auxiliary organismal functions such as pathogenesis, antibiotic resistance, symbiosis, and specialized metabolism [38]. Despite some parallels, the mechanisms by which bacterial genomic islands and unstable gene islands in animals arise are different: clustering in bacteria is a consequence of horizontal transfer of contiguous blocks of DNA, whereas clustering in animals is restricted to closely-related genes and arises from local gene duplication events without horizontal gene transfer.

Materials and Methods

Sequence datasets. Sequence data were downloaded from ENSEMBL [39] release 40, which contained the following genome builds: *H. sapiens* NCBI36, *M. mulatta* MMUL1, *B. taurus* Btau2.0, *C. familiaris* BROADD1, *M. musculus* NCBI36, *R. norvegicus* RGSC3.4, *G. gallus* WASHUC1, *X. tropicalis* JGI4.1, *D. rerio* ZFISH6, and *T. rubripes* FUGU4.40. Data for *P. troglodytes* were not included because the genes are too similar to human to be informative. Data for *Monodelphis domestica* and *Tetraodon nigroviridis* and several low coverage genomes were considered but rejected because current gene prediction sets appeared to be too incomplete or inaccurate (unpublished data). Starting with a hand-collected set of human CYP450 sequences, all known isoforms of human CYP450 proteins were collected from a protein BLAST (BLASTP) search of all human proteins. Many genes had multiple transcript and protein forms. Each such case was analyzed to identify a single protein isoform that appeared full length and most family-typical, based on alignment with related CYP450 proteins. This analysis resulted in a set of 60 reference human proteins that were full length or nearly full length, including three possible pseudogenes. Proteins from other species were collected from BLASTP searches with these 60 human proteins as query. All isoforms were kept initially, and for each gene the isoform with the highest BLASTP score against a human protein was kept for further analysis (for the most part, this approach discarded obviously incomplete proteins). Subsequent tree analysis of these proteins indicated the surprising absence of a few CYP450 proteins with endogenous substrates. Some of these were found to be present (but unpredicted) in the cognate genomes by TBLASTN searches (Table S1). Other genes are probably present in unassembled sequence; the number of missing genes and genome sequence coverage in the mammalian genomes is summarized in Table S2.

Protein trees. After discarding a few proteins that were missing more than 30% of the expected CYP450 protein or clearly aligned badly, the remaining set of 628 proteins were aligned with ClustalW (default settings except gap clustering turned off) [40]. A complete list of these protein sequences is found in Dataset S1 with ENSEMBL gene identifiers [39]. After removal of aligned columns with more than 30% gap residues, the resulting multiple alignment had 417 sites, at least 70% of which were present in any given sequence. A maximum-likelihood tree was constructed from this alignment with PhyML (JTT matrix, four rate categories, gamma shape parameter 1.0) [41]. Tree nodes were rotated, colored, and displayed with *Bonsai* [42]. Additional annotations were added to the tree image for Figure S1. The complete tree was subjected to 200 bootstrap repeats with PhyML with the same settings, which required ten days of central processing unit time. Subtrees of specific interest were tested with 1,000 bootstrap repeats (Figure S2).

Codon analysis for positive selection. Multiple sets of five to 15 closely related primate and rodent genes were selected from the tree

in Figure S1. For each set, a protein alignment was made using ClustalW (default settings except gap clustering turned off) [40]. This protein alignment was used to generate the corresponding codon alignment and to construct a maximum-likelihood protein tree with PhyML [41]. The tree and codon alignment were analyzed with CodeML from PAML 3.14 [14], using models 7 and 8, with three starting d_N/d_S (ω) values for model 8 to avoid local optima. The neutral model 7 assumes a β -distribution of 10 d_N/d_S ratio classes constrained to lie between 0 and 1.0, whereas the selection model 8 permits one additional d_N/d_S ratio class without constraint. In order to minimize the effects of gene prediction and alignment errors, aligned columns with a gap in any sequence were excluded from analysis (“cleandata” option in CodeML). The transition/transversion ratio (κ) was estimated by CodeML. Statistical significance was assessed using a χ -square test on twice the difference in log-likelihood values (Δ ML) for models 7 and 8 with two degrees of freedom, a statistic shown to be conservative in simulations [43]. Assignment of specific sites under likely positive selection was based on the Bayes-Empirical-Bayes test implemented in CodeML. In Figures S6 and S7, marked sites are subject to positive selection with $p > 0.9$. Extensive gene conversion can exaggerate signals of positive selection, but direct analysis of gene conversions among the tested genes suggests that they are unlikely to confound the analysis. GENECONV [44] was used to test for possible gene conversion events in the complete genomic regions of all the mouse and human genes analyzed for positive selection (unpublished data). In all cases except one, detected regions of conversion affected only a few genes in <10% of their coding region, suggesting that they will not significantly affect the signal of positive selection. The exception was among mouse genes in the *CYP2D* family, but even in this case a minority of genes had evidence of gene conversion and none were affected in more than 20% of their coding region.

Structure analysis. The significance of the association of positive selection sites to bound warfarin ligand in the *CYP2C9* structure was assessed as follows. The mean physical distance of the alpha backbone carbon atom to a reference warfarin atom was measured for the nine amino acid residues under positive selection. Warfarin atom C10 (carbon 10, atom 7439 of Protein Data Base structure 10G5) was used as the reference; it was selected as centrally located in the binding pocket/access channel based on visual inspection of the three-dimensional protein structure. The expected mean distance for randomly located amino acid residues was determined by repeatedly choosing nine random amino acid residues and computing the mean distance from their alpha backbone carbon atoms to the reference warfarin atom. Mean distances for 10,000 simulation repeats fitted a normal distribution (Figure S8), and all but two had mean distances larger than the mean distance for positively selected sites.

Supporting Information

Dataset S1. Protein Sequences

Found at 10.1371/journal.pgen.0030067.sd001 (315 KB XLS).

Figure S1. Maximum-Likelihood Tree of 628 Vertebrate CYP450 Proteins

Filled black square: human protein with known endogenous substrate. Filled blue circle: human protein with known xenobiotic substrates. Red asterisks indicate positive selection was detected; tests for positive selection were performed only on primate and rodent genes. Other markings indicate enzyme activity, gene clusters in human (blue) and mouse (pink), family groups (black), and specific groups tested where positive selection was not found (green). References for specific endogenous-substrate gene functions are as follows. CYP26C1 (retinoic acid catabolizing enzyme) [45]; CYP26B1 (retinoic acid catabolizing enzyme) and CYP26A1 (retinoic acid 4-hydroxylase) [46]; CYP51A1 (sterol 14- α -demethylase) [47]; CYP39A1 (oxysterol 7- α -hydroxylase) [48]; CYP8A1 (prostaglandin synthase) [49]; CYP8B1 (sterol 12- α -hydroxylase) [50]; CYP7B1 (oxysterol 7- α -hydroxylase) [51,52]; CYP7A1 (cholesterol 7- α -hydroxylase) [53,54]; CYP19A1 (estrogen synthetase or aromatase) [55]; CYP46A1 (cholesterol 24-hydroxylase) [56]; CYP5A1 (thromboxane-A synthase) [57,58]; CYP4V2 (fatty-acid metabolism) [59]; CYP24A1 (vitamin D3 24-hydroxylase) [60]; CYP11A1 (cholesterol side-chain cleavage enzyme) [61]; CYP11B2 (aldosterone synthase) and CYP11B1 (steroid 11- β -hydroxylase) [62]; CYP27A1 (sterol 26-hydroxylase) [63]; CYP27B1 (25-hydroxyvitamin D3 1- α -hydroxylase) [64]; CYP21A2 (steroid 21-hydroxylase) [65]; CYP17A1 (steroid 17- α -monooxygenase) [66]; CYP1B1 (steroid catabolism) [67]. Evidence for xenobiotic activity

of CYP450 enzymes is often scattered among many papers; the markings here are based primarily on two recent reviews [6,7].

Found at 10.1371/journal.pgen.0030067.sg001 (2.6 MB JPG).

Figure S2. Bootstrap Tests on Subtrees

A total of four sets of proteins that form clades on the tree in Figure S1 were tested by more extensive bootstrap analysis. The four sets are from the CYP3A, CYP4F, CYP2C, and CYP2A/B families. In each case, the proteins from the family were aligned and subjected to 1,000 bootstrap repeats as described in Materials and Methods. Bootstrap values shown (percentage branch support) are those that are relevant to confidence of species-specific gene duplications. In a separate analysis, two outgroup proteins were added to the set for rooting.

Found at 10.1371/journal.pgen.0030067.sg002 (1.0 MB JPG).

Figure S3. Relationship of Genes Within and among the Mouse and Human CYP2C Unstable Gene Islands

The figure is derived from snapshots on the University of California Santa Cruz Genome Browser (<http://genome.ucsc.edu>) in October 2006. Orthologous gene pairs flanking the unstable *CYP450* genes are marked with green lines, which include one pair of *CYP450* genes (human *CYP2C18* and mouse *Cyp2c55* at the left end of the unstable island). ENSEMBL gene identifiers are shown for two genes without formal mouse gene names. The unstable genes in the island include three human *CYP450* genes and 12 mouse *Cyp450* genes. A protein tree relating all the genes in these clusters is shown in Figure S2 (bottom left).

Found at 10.1371/journal.pgen.0030067.sg003 (599 KB JPG).

Figure S4. Relationship of Genes within and among the Mouse and Human CYP3A Unstable Gene Islands

See Figure S3 legend description. This gene cluster underwent a single rearrangement on the mouse lineage that split the mouse cluster into two parts, which are shown separately aligned to the human cluster. The unstable genes in the island include four human *CYP450* genes and nine mouse *Cyp450* genes. Positions of two ENSEMBL predicted mouse *Cyp450* genes that did not appear on the UCSC browser were added by hand.

Found at 10.1371/journal.pgen.0030067.sg004 (575 KB JPG).

Figure S5. Orthology and Synteny among Human and Macaque CYP3A Genes

Human genes are labeled in blue and macaque genes are labeled in green. Both species have four *CYP3A* genes that are present in a single gene cluster, and gene pairs marked syntenic occupy equivalent positions in their respective clusters. The tree is inconsistent with significant concerted evolution subsequent to separation of the two lineages. The proteins encoded by each gene were aligned and subjected to maximum-likelihood tree analysis with 1,000 bootstrap repeats as described in Materials and Methods. Bootstrap values shown (percentage of branch support) are those relevant to determining orthology. A total of two of the macaque genes were not predicted in the current assembly (MMUL1); they were identified by TBLASTN searches with the human protein as query. For each, a gene prediction was hand assembled, and the protein is labeled with chromosome and position information in place of a gene name.

Found at 10.1371/journal.pgen.0030067.sg005 (260 KB JPG).

Figure S6. Positive Selection Sites Mapped onto a Rodent CYP2D Alignment

Full length alignments of proteins from the 12 rodent genes with the strongest signal of positive selection and two closely related human proteins with crystal structures. Blue shading is proportional to sum-of-pairs conservation for each residue with respect to its aligned column. Major alpha helices are marked above the alignment for context. Sites with significant evidence of positive selection ($p > 0.9$) are marked with red asterisks and numbered below the alignment; numbers correspond with those in Figure S7. A total of three sites of indels among the rodent genes are marked with a red D. Residues that line the substrate-binding cavity in the crystal structures are marked with an orange S, and residues that form the substrate-entry channel are marked with an orange C. Heme-binding residues are marked with a green H. The binding site, entry channel, and heme-binding assignments are based solely on the published description [17]. Positive selection site 1 is also located close to the substrate-entry channel (Figure S7).

Found at 10.1371/journal.pgen.0030067.sg006 (2.3 MB JPG).

Figure S7. Positive Selection Sites Mapped onto the CYP2C9 Structure

Red space-filled residues align with sites of positive selection in the set of proteins in Figure S6. The alignment positions of two indel sites are marked on the backbone with red arrowheads; the third indel site is hidden behind the residue marked 9 on a loop that forms part of the substrate-binding cavity. The bound warfarin ligand is shown in purple, and the heme is shown in brown. A second probable substrate-binding site is located between the warfarin and the heme but is unoccupied in this structure [17]. Alpha helices are shown by gray cylinders, and their lettering corresponds to that in Figure S6. The olive section of backbone is membrane associated.

Found at 10.1371/journal.pgen.0030067.sg007 (600 KB JPG).

Figure S8. Distribution of Mean Distances from Random Sampling of CYP2C9

The mean distance to the reference warfarin atom C10 from 10,000 samples of nine random residues from the CYP2C9 structure is shown. The sampled distances were put into bins of 1 Angstrom for graphing. The black arrow indicates the mean distance for the nine sites of positive selection to the same warfarin atom (14.4 Angstroms).

Found at 10.1371/journal.pgen.0030067.sg008 (164 KB JPG).

Table S1. Results of TBLASTN Fishing for Unpredicted Genes

For each case of an expected mammalian gene that was not in the predicted set for a genome, the protein from another species was used as query for a TBLASTN search of the genome (–m 9 option, no filtering, E-value cutoff 10). For cases where no good match was found, the table has an entry giving only the query and the result (in most cases there were no matches with an E-value $< 10^{-3}$). For cases where a strong candidate gene was found, the TBLASTN output is shown, with unrelated matches removed and sorted by query position.

Found at 10.1371/journal.pgen.0030067.st001 (16 KB XLS).

Table S2. Summary of Probable Missing Stable Genes

Summary of gene content among mammals for the 20 most stable groups of CYP450 enzymes with known endogenous substrates (all proteins marked with a black square on Figure S1, except for the pair of corticosteroid enzymes CYP11B1 and CYP11B2). A total of 20 genes are expected for each genome, but one to three genes were missing from predicted gene sets for all the genomes except human and mouse. Column 3 summarizes the results from Table S1 (putative genes found by TBLASTN searches), and column 4 indicates the number of genes that remain missing. Column 5 summarizes the estimated single-copy sequence coverage for each genome, based on NCBI genome summaries. Not all of this sequence is present in the assemblies used for TBLASTN searching.

Found at 10.1371/journal.pgen.0030067.st002 (14 KB XLS).

Table S3. Summary of Statistics for Maximum-Likelihood Analysis of Positive Selection

In each column the CYP family, coding sequence file name, number, and mean length of sequences are given, followed by analysis results relevant to inferring positive selection. For cases in which the difference in maximum-likelihood for models 7 and 8 was large, a p -value was computed (see Materials and Methods) with Bonferroni correction. Total tree length is a significant parameter for sensitivity [43]. The “model 8 11th class omega” value is the value for the free d_N/d_S class in model 8, which was much greater than 1.0 in all cases of high significance.

Found at 10.1371/journal.pgen.0030067.st003 (18 KB XLS).

Accession Numbers

The Protein Data Base (<http://www.pdb.org>) number for the structure discussed in this paper is 1OG5.

Acknowledgments

I thank Barbara Trask, Nathan Clarke, Joshua McElwee, and members of the Thomas lab for helpful input to this work.

Funding. The author received no specific funding for this study.

Competing interests. The author has declared that no competing interests exist.

References

- Hughes AL, Nei M (1989) Evolution of the major histocompatibility complex: Independent origin of nonclassical class I genes in different groups of mammals. *Mol Biol Evol* 6: 559–579.
- Ota T, Nei M (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol Biol Evol* 11: 469–482.
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155.
- Hurley I, Hale ME, Prince VE (2005) Duplication events and the evolution of segmental identity. *Evol Dev* 7: 556–567.
- Lynch M (2002) Genomics. Gene duplication and evolution. *Science* 297: 945–947.
- Guengerich FP (2003) Cytochromes P450, drugs, and diseases. *Mol Interv* 3: 194–204.
- Ding X, Kaminsky LS (2003) Human extrahepatic cytochromes P450: Function in xenobiotic metabolism and tissue-selective chemical toxicity in the respiratory and gastrointestinal tracts. *Annu Rev Pharmacol Toxicol* 43: 149–173.
- Bernhardt R (2006) Cytochromes P450 as versatile biocatalysts. *J Biotechnol* 124: 128–145.
- Guengerich FP, Shimada T (1991) Oxidation of toxic and carcinogenic chemicals by human cytochrome P-450 enzymes. *Chem Res Toxicol* 4: 391–407.
- Woods IG, Wilson C, Friedlander B, Chang P, Reyes DK, et al. (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* 15: 1307–1314.
- Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946–957.
- Colombo L, Dalla Valle L, Fiore C, Armanini D, Belvedere P (2006) Aldosterone and the conquest of land. *J Endocrinol Invest* 29: 373–379.
- Goldstone HM, Stegeman JJ (2006) A revised evolutionary history of the CYP1A subfamily: Gene duplication, gene conversion, and positive selection. *J Mol Evol* 62: 708–717.
- Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556.
- Thomas JH, Kelley JL, Robertson HM, Ly K, Swanson WJ (2005) Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc Natl Acad Sci U S A* 102: 4476–4481.
- Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, et al. (2006) Crystal structure of human cytochrome P450 2D6. *J Biol Chem* 281: 7614–7622.
- Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovic D, et al. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* 424: 464–468.
- Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19: 950–958.
- de Bono B, Madera M, Chothia C (2004) VH gene segments in the mouse and human genomes. *J Mol Biol* 342: 131–143.
- Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci U S A* 94: 7799–7806.
- Su C, Jakobsen I, Gu X, Nei M (1999) Diversity and evolution of T-cell receptor variable region genes in mammals and birds. *Immunogenetics* 50: 301–308.
- Niimura Y, Nei M (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene* 346: 23–28.
- Chen N, Pai S, Zhao Z, Mah A, Newbury R, et al. (2005) Identification of a nematode chemosensory gene family. *Proc Natl Acad Sci U S A* 102: 146–151.
- Hardison R (1998) Hemoglobins from bacteria to man: Evolution of different patterns of gene expression. *J Exp Biol* 201: 1099–1117.
- Aguileta G, Bielawski JP, Yang Z (2004) Gene conversion and functional divergence in the beta-globin gene family. *J Mol Evol* 59: 177–189.
- Parry JW, Carleton KL, Spady T, Carboo A, Hunt DM, et al. (2005) Mix and match color vision: Tuning spectral sensitivity by differential opsin gene expression in Lake Malawi cichlids. *Curr Biol* 15: 1734–1739.
- Spady TC, Parry JW, Robinson PR, Hunt DM, Bowmaker JK, et al. (2006) Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Mol Biol Evol* 23: 1538–1547.
- Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res* 16: 669–677.
- Thomas JH (2006) Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res* 16: 1017–1030.
- Genomics Institute of the Novartis Research Foundation (2003) GNF SymAtlas, version 1.2.4 [web application and database]. Available: <http://symatlas.gnf.org/SymAtlas>. Accessed 10 March 2007.
- Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, et al. (2006) Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res* 16: 584–594.
- Niimura Y, Nei M (2003) Evolution of olfactory receptor genes in the human genome. *Proc Natl Acad Sci U S A* 100: 12235–12240.
- Newman T, Trask BJ (2003) Complex evolution of 7E olfactory receptor genes in segmental duplications. *Genome Res* 13: 781–793.
- Bailey JA, Eichler EE (2006) Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat Rev Genet* 7: 552–564.
- Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, et al. (2005) A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* 15: 1344–1356.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. *Science* 297: 1003–1007.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2: 414–424.
- Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, et al. (2004) Ensembl 2004. *Nucleic Acids Res* 32 Database issue: D468–D470.
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680.
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696–704.
- Thomas JH (2004) Bonsai 1.1.4 download, March, 2004. Available: <http://depts.washington.edu/jtlab/software/softwareIndex.html>. Accessed 10 March 2007.
- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18: 1585–1592.
- Sawyer S (2007) GENECONV molecular biology computer program, version 1.81. Available: <http://www.math.wustl.edu/~sawyer/geneconv>. Accessed 10 March 2007.
- Taimi M, Helvig C, Wisniewski J, Ramshaw H, White J, et al. (2004) A novel human cytochrome P450, CYP26C1, involved in metabolism of 9-cis and all-trans isomers of retinoic acid. *J Biol Chem* 279: 77–85.
- White JA, Beckett-Jones B, Guo YD, Dilworth FJ, Bonasoro J, et al. (1997) cDNA cloning of human retinoic acid-metabolizing enzyme (hP450RAI) identifies a novel family of cytochromes P450. *J Biol Chem* 272: 18538–18541.
- Fischer RT, Stam SH, Johnson PR, Ko SS, Magolda RL, et al. (1989) Mechanistic studies of lanosterol 14 alpha-methyl demethylase: Substrate requirements for the component reactions catalyzed by a single cytochrome P-450 isozyme. *J Lipid Res* 30: 1621–1632.
- Li-Hawkins J, Lund EG, Bronson AD, Russell DW (2000) Expression cloning of an oxysterol 7alpha-hydroxylase selective for 24-hydroxycholesterol. *J Biol Chem* 275: 16543–16549.
- Miyata A, Hara S, Yokoyama C, Inoue H, Ullrich V, et al. (1994) Molecular cloning and expression of human prostacyclin synthase. *Biochem Biophys Res Commun* 200: 1728–1734.
- Gavels M, Olin M, Chowdhary BP, Raudsepp T, Andersson U, et al. (1999) Structure and chromosomal assignment of the sterol 12alpha-hydroxylase gene (*CYP8B1*) in human and mouse: Eukaryotic cytochrome P-450 gene devoid of introns. *Genomics* 56: 184–196.
- Schwarz M, Lund EG, Lathe R, Bjorkhem I, Russell DW (1997) Identification and characterization of a mouse oxysterol 7alpha-hydroxylase cDNA. *J Biol Chem* 272: 23995–24001.
- Rose KA, Stapleton G, Dott K, Kienny MP, Best R, et al. (1997) Cyp7b, a novel brain cytochrome P450, catalyzes the synthesis of neurosteroids 7alpha-hydroxy dehydroepiandrosterone and 7alpha-hydroxy pregnenolone. *Proc Natl Acad Sci U S A* 94: 4925–4930.
- Jelinek DF, Andersson S, Slaughter CA, Russell DW (1990) Cloning and regulation of cholesterol 7 alpha-hydroxylase, the rate-limiting enzyme in bile acid biosynthesis. *J Biol Chem* 265: 8190–8197.
- Noshiro M, Nishimoto M, Morohashi K, Okuda K (1989) Molecular cloning of cDNA for cholesterol 7 alpha-hydroxylase from rat liver microsomes. Nucleotide sequence and expression. *FEBS Lett* 257: 97–100.
- Evans CT, Ledesma DB, Schulz TZ, Simpson ER, Mendelson CR (1986) Isolation and characterization of a complementary DNA specific for human aromatase-system cytochrome P-450 mRNA. *Proc Natl Acad Sci U S A* 83: 6387–6391.
- Lund EG, Guileyardo JM, Russell DW (1999) cDNA cloning of cholesterol 24-hydroxylase, a mediator of cholesterol homeostasis in the brain. *Proc Natl Acad Sci U S A* 96: 7238–7243.
- Yokoyama C, Miyata A, Ihara H, Ullrich V, Tanabe T (1991) Molecular cloning of human platelet thromboxane A synthase. *Biochem Biophys Res Commun* 178: 1479–1484.
- Wang LH, Ohashi K, Wu KK (1991) Isolation of partial complementary DNA encoding human thromboxane synthase. *Biochem Biophys Res Commun* 177: 286–291.
- Li A, Jiao X, Munier FL, Schorderet DF, Yao W, et al. (2004) Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. *Am J Hum Genet* 74: 817–826.
- Ohayama Y, Noshiro M, Okuda K (1991) Cloning and expression of cDNA encoding 25-hydroxyvitamin D3 24-hydroxylase. *FEBS Lett* 278: 195–198.
- Chung BC, Matteson KJ, Voutilainen R, Mohandas TK, Miller WL (1986) Human cholesterol side-chain cleavage enzyme, P450sc: cDNA cloning,

- assignment of the gene to chromosome 15, and expression in the placenta. *Proc Natl Acad Sci U S A* 83: 8962–8966.
62. Nonaka Y, Okamoto M (1991) Functional expression of the cDNAs encoding rat 11 beta-hydroxylase [cytochrome P450(11 beta)] and aldosterone synthase [cytochrome P450(11 beta, aldo)]. *Eur J Biochem* 202: 897–902.
 63. Andersson S, Davis DL, Dahlback H, Jornvall H, Russell DW (1989) Cloning, structure, and expression of the mitochondrial cytochrome P-450 sterol 26-hydroxylase, a bile acid biosynthetic enzyme. *J Biol Chem* 264: 8222–8229.
 64. Takeyama K, Kitanaka S, Sato T, Kobori M, Yanagisawa J, et al. (1997) 25-Hydroxyvitamin D3 1alpha-hydroxylase and vitamin D synthesis. *Science* 277: 1827–1830.
 65. White PC, Grossberger D, Onufer BJ, Chaplin DD, New MI, et al. (1985) Two genes encoding steroid 21-hydroxylase are located near the genes encoding the fourth component of complement in man. *Proc Natl Acad Sci U S A* 82: 1089–1093.
 66. Chung BC, Picado-Leonard J, Haniu M, Bienkowski M, Hall PF, et al. (1987) Cytochrome P450c17 (steroid 17 alpha-hydroxylase/17,20 lyase): Cloning of human adrenal and testis cDNAs indicates the same gene is expressed in both tissues. *Proc Natl Acad Sci U S A* 84: 407–411.
 67. Sutter TR, Tang YM, Hayes CL, Wo YY, Jabs EW, et al. (1994) Complete cDNA sequence of a human dioxin-inducible mRNA identifies a new gene subfamily of cytochrome P450 that maps to chromosome 2. *J Biol Chem* 269: 13092–13099.