A revised evolutionary history of the CYP1A subfamily: Gene duplication, gene conversion, and positive selection

Heather M. H. Goldstone[1,2] and John J. Stegeman[1]

[1] Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

[2] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

Correspondence:
Heather M. H. Goldstone
Marine Biological Laboratory
7 MBL Street
Woods Hole MA 02543
Phone: (508) 548-3705 x6623
Fax: (508) 457-4727
E-mail: hhandley@alum.mit.edu

Running header:
Recombination and selection in tetrapod CYP1As

Key words:
Cytochrome P450, gene conversion, gene duplication, chicken, mammalian

Abbreviations:
CYP1A=cytochrome P450 1A

ABSTRACT

Members of cytochrome P450 subfamily 1A (CYP1As) are involved in detoxification and bioactivation of common environmental pollutants. Understanding the functional evolution of these genes is essential to predicting and interpreting species differences in sensitivity to toxicity by such chemicals. The CYP1A gene subfamily comprises a single ancestral representative in most fish species and two paralogs in higher vertebrates, including birds and mammals. Phylogenetic analysis of complete coding sequences suggests that mammalian and bird paralog pairs (CYP1A1/2 and CYP1A4/5, respectively) are the result of independent gene duplication events. However, comparison of vertebrate genome sequences revealed that CYP1A genes lie within an extended region of conserved fine-scale synteny, suggesting that avian and mammalian CYP1A paralogs share a common genomic history. Algorithms designed to detect recombination between nucleotide sequences indicate that gene conversion has homogenized most of the length of the chicken CYP1A genes, as well as the 5' end of mammalian CYP1As. Together, these data indicate that avian and mammalian CYP1A paralog pairs resulted from a single gene duplication event and that extensive gene conversion is responsible for the exceptionally high degree of sequence similarity between CYP1A4 and CYP1A5. Elevated non-synonymous/synonymous substitution ratios within a putatively unconverted stretch of ~250 bp suggests that positive selection may have reduced the effective rate of gene conversion in this region, which contains two substrate recognition sites. This work significantly alters our understanding of functional evolution in the CYP1A subfamily, suggesting that gene conversion and positive selection have been the dominant processes of sequence evolution.

INTRODUCTION

The cytochrome P450 (CYP) superfamily comprises more than 5,000 genes encoding heme-thiolate enzymes that catalyze the oxidative metabolism of a vast array of organic compounds. Members of cytochrome P450 gene family 1 (CYP1s) have a broad affinity for polycyclic, often halogenated, aromatic hydrocarbons, as well as aromatic amines, and some endogenous substrates. CYP1s play important roles in both mediating and mitigating the biological effects of these chemicals and can determine susceptibility to toxicity or disease (Elskus *et al.*, 1999; reviewed in Gonzalez and Kimura, 2003; Teraoka *et al.*, 2003). Thus, understanding species differences in CYP1 complement and function is of significant biomedical and toxicological import.

The CYP1 gene family consists of three subfamilies – the CYP1As, CYP1Bs and CYP1Cs (Godard *et al.*, 2005). The CYP1A subfamily appears to have originated early in the vertebrate lineage. Fish generally possess a single CYP1A gene (e.g. Morrison *et al.*, 1995; Morrison *et al.*, 1998); eels and salmonids are notable exceptions (Rabergh *et al.*, 2000; Mahata *et al.*, 2003). Mammals, in contrast, generally possess two paralogous CYP1A genes, CYP1A1 and CYP1A2 (e.g. Kimura *et al.*, 1984; Okino *et al.*, 1985; Quattrochi *et al.*, 1985). Fish CYP1As share significant sequence similarity with both CYP1A1s and CYP1A2s (Morrison *et al.*, 1995) and display a combination of catalytic functions characteristic of the mammalian isoforms (e.g. Gorman *et al.*, 1998). However, fish CYP1As are considered more CYP1A1-like on the basis of slightly higher levels of pairwise sequence identity and similarities in patterns of gene expression.

Birds also possess genes for two CYP1A isoforms (Rifkind *et al.*, 1994; Gilday *et al.*, 1996c), and there is evidence from biochemical studies that turtles express more than one

CYP1A isoform (Yawetz *et al.*, 1998). Phylogenetic analyses of nucleic acid and amino acid sequences show that chicken CYP1A4 and CYP1A5 cluster together, separately from the mammalian CYP1A1/1A2 clade (Gilday *et al.*, 1996b). This has been interpreted to mean that CYP1A4 and CYP1A5 resulted from an independent gene duplication that occurred subsequent to the divergence of the avian and mammalian lineages. However, the catalytic specificities of chicken CYP1A4 and CYP1A5 bear a distinct resemblance to mammalian CYP1A1 and CYP1A2, respectively. Ethoxyresorufin-O-deethylase (EROD) activity by CYP1A1 and CYP1A4 is approximately ten-fold greater than by CYP1A2 and CYP1A5, whereas CYP1A2 and CYP1A5 are primarily responsible for uroporphyrinogen oxidation (UROX) (Rifkind *et al.*, 1994; Sinclair *et al.*, 1998).

Establishing orthologous relationships among CYP1As, or indeed among any set of CYP genes, is crucial to relating results from studies using different species, including model organisms. However, the evolutionary history of the CYP superfamily appears to be extremely complex. Gene and genome duplication, gene amplification and conversion, gene structure rearrangements (gain or loss of introns/exons), gene loss, horizontal gene transfer, and convergent evolution have all been implicated in CYP evolution (reviewed by Werck-Reichhart and Feyereisen, 2000). As a result, ortholog/paralog assignment can be extremely difficult.

The available data regarding chicken CYP1As can be reconciled in either of two ways – independent duplications followed by convergent evolution, or a single gene duplication event followed by concerted evolution. The hypothesis suggested by phylogenetic analyses is that CYP1A4 and CYP1A5 are the result of an independent duplication but have been subject to the same selective pressures as CYP1A1 and CYP1A2 and, thus, evolved similar functional characteristics (convergent functional evolution). Alternatively, a single gene duplication event

prior to avian-mammalian divergence could have generated the ancestral CYP1A1/4 and CYP1A2/5 genes, and greater similarity between CYP1A4 and CYP1A5 than between avian and mammalian orthologs could be the result of concerted evolution.

Concerted evolution via interlocus gene conversion is increasingly recognized as a major feature of evolution of small multigene families (e.g. Michelson and Orkin, 1983; Jouvin-Marche *et al.*, 1986; Rudikoff *et al.*, 1992; Aguileta *et al.*, 2004; e.g. Teshima and Innan, 2004). Gene conversion can be functionally defined as the non-reciprocal, or asymmetric, transfer of material from one strand or region of DNA to another. The precise mechanisms of interlocus gene conversion remain unclear, but the consequence is the homogenization of variation between paralogous gene copies (Teshima and Innan, 2004). As a result, sequences of converted paralogs may come to resemble one another more than they do orthologous sequences in other species.

Here we present several lines of evidence in support of the single duplication/gene conversion hypothesis for the CYP1A4/1A5 and the CYP1A1/1A2 paralog pairs. An extended region of conserved fine-scale synteny encompasses mammalian, chicken, and to a lesser extent, amphibian and fish CYP1As. Based on this synteny, we conclude that chicken and mammalian CYP1As are the result of the same tandem inverted gene duplication event. Spatial heterogeneity in phylogenetic signal strongly suggests that all examined CYP1A genes have undergone partial gene conversion, and several gene conversion events are supported by at least one additional recombination detection algorithm. However, positive selection appears to have excluded gene conversion in a region of ~250 bp, which we suggest may be responsible for isoform-specific catalytic activities. We further suggest that gene conversion may be a common feature of tandemly duplicated CYP genes and obfuscates CYP relationships.

METHODS

Genomic DNA sequences and fine-scale synteny data were retrieved from Ensembl v30.1 (released 22 March 2005). Manual annotation of chicken CYP1A4 and CYP1A5 genes was carried out using GCG Wisconsin Package (Accelrys, Inc.). Multi-sequence alignments for use in recombination detection algorithms (below) were generated in ClustalX (Thompson *et al.*, 1997). Poorly aligned regions at the 5' and 3' termini were removed to yield a final alignment of 1498 bp (see Supplementary Material). Nucleotide positions refer to aligned positions, not the absolute position within an individual sequence.

The hypothesis of recombination between putative paralogs was tested using CODOUBLE (Drouin *et al.*, 1999), which is an expanded implementation of the algorithm introduced by Balding, Nichols, and Hunt (Balding *et al.*, 1992). All possible pairwise comparisons (e.g. chicken vs. human, mouse vs. human, etc.) were performed and p-values were subjected to Bonferroni correction for multiple comparisons.

Spatial variation in phylogenetic signal was detected with SlidingBayes (Paraskevis *et al.*, 2005), using a window size of 100bp and step size of 10bp. For each window, 120,000 generations of Metropolis-coupled Markov chain Monte Carlo (MC^3) simulation was run using a general time-reversible model of nucleotide substitution with substitution rates modeled by a γ distribution with 8 rate categories. Convergence of MC^3 simulations for selected windows was determined on the basis of plateaus in both the log likelihood score of the best tree and the posterior probability of each split within the tree (AWTY, http://ceb.csit.fsu.edu/awty). Trees were sampled every 50 generations. Burn-in was defined at 20,000 generations; remaining trees were used to determine the Bayesian posterior probabilities of selected nodes.

The RDP2 software package (Martin *et al.*, 2005) was used to run simultaneous analyses with five recombination detection algorithms – RDP, GeneConv, MaxChi, Bootscan, Chimaera, and Reticulate. RDP was run with default settings modified to use internal and external reference sequences. Default settings were used for GeneConv analysis. MaxChi and Chimaera default settings were modified to use variable window sizes with 10% variable sites and to perform 1000 permutations. Bootscan was run using a window size of 100bp and a step size of 10bp; 1000 bootstrap replicates were performed with cutoff percentage set to 95% and the bootstrap value used as p-value. P-values from individual algorithms were Bonferroni corrected to account for both multiple testing within each method and for multiple tests by different methods. Only events with corrected p-values < 0.01 were considered. Compatibility matrices were generated by Reticulate (a) using all informative sites (sites with multiple states included) or (b) using only binary sites. 1000 randomized matrices were generated for reference.

PAML (Yang, 1997) was used to test for evidence of positive selection acting on tetrapod CYP1A genes. We applied the F3x4 codon model and allowed for maximum likelihood estimation of κ (transition:transversion ratio) and ω (non-synonymous:synonymous substitution ratio). We used the likelihood ratio test to compare the fit of a homogeneous model (single ω for all sites in all branches) to that of a branch model in which all branches were allowed independent ω values. P-values were based on critical values for a χ^2 distribution with 12 degrees of freedom.

RESULTS

Conserved fine-scale synteny between vertebrate CYP1A loci

BLAST searches of the chicken genome using previously published transcript sequences (GenBank Reference Sequences NM_205147 & NM_205146) (Gilday *et al.*, 1996c) revealed

that CYP1A4 and CYP1A5 are arranged head-to-head on chromosome 10. This arrangement does not appear to be a genome assembly artifact, as there is a clear gap between the genes and comparisons with known transcript sequences show no evidence of chimeric misassembly. The CYP1A5 transcriptional start site is located at nucleotide position 15,790 on the minus strand of chromosome 10. The CYP1A4 gene begins ≥1 kb upstream of CYP1A5 on the opposite (i.e. plus) strand. Mammalian CYP1A1 and CYP1A2 genes are also arranged in head-to-head tandem inverted configurations.

Chicken chromosome 10 shares large tracts of conserved synteny with human chromosome 15 (Crooijmans *et al.*, 2001), which houses the CYP1A genes. Thus, we undertook a detailed investigation of shared synteny between the chicken CYP1A4/1A5 locus and the corresponding regions in mammals for which complete genome sequence assemblies are available (i.e., human, chimpanzee, mouse, rat, and dog). Comparison between chicken chromosome 10 and human chromosome 15q23 sequences revealed an extended region of conserved fine-scale synteny. Gene order and orientation were conserved between human and chicken for 9 out of 10 genes in the first 130 kb of chicken chromosome 10 (Figure 1). Chicken cytoplasmic actin type 5 (*act5*) is homologous to human gamma actin (actg), found on chromosome 17. The corresponding regions in mouse, rat, dog, and chimpanzee also manifest gene order and orientation identical to human (not shown). Patchier synteny was observed out to 180 kb (not shown).

The CYP1A locus and surrounding region is significantly larger in mammals than in chicken. The region encompassing CYP11A1 through CSK spans approximately 450 kb in human, but less than 135 kb in chicken. Differences in the lengths of both intergenic and intronic sequences associated with CYP1A genes are evident, but the difference is by far the most pronounced in intergenic space. A sequencing gap obscures the 5' end (~700 bp) of the CYP1A4 transcript,

making it impossible to determine precisely the distance from transcriptional start to transcriptional start. However, 1 kb of intergenic sequence is available currently. Assuming that the estimated gap size is within 5-fold of reality, the CYP1A4-CYP1A5 intergenic distance could be as great as ~3 kb. In contrast, CYP1A1 and CYP1A2 are separated by 13-17 kb in rodents and 21-24 kb in primates. Intronic sequence length is also greater in mammalian than in chicken CYP1A genes. Human CYP1A1 and CYP1A2 are 5986 bp and 6248 bp in length, respectively, as opposed to approximately 3 kb for each of the chicken CYP1A genes.

Early diverging vertebrates also demonstrate synteny with the mammalian CYP1A locus and the region downstream of CYP1A2 (Figure 1). The gene series CYP1A-CSK-LMAN1L-ULK3 is found on an unidentified chromosome in the pufferfish (*Tetraodon nigridis*) genome. Gene orientation in this region corresponds to mammalian CYP1A2-CSK-LMAN1L-ULK3. A similar region is found at the end of scaffold 287 from the preliminary assembly of the frog (*Xenopus tropicalis*) genome sequence. In frog, the region of conserved gene order and orientation extends an additional 4 genes beyond ULK3 (not shown). Notably, the orientation of the frog CYP1A gene is like that of CYP1A4 and mammalian CYP1A1 genes, but opposite of that in pufferfish.

Reconciling synteny data and phylogenetic reconstructions

These genome comparisons strongly suggested that avian and mammalian CYP1A paralogs were the result of a single tandem inverted gene duplication event. We used CODOUBLE to test whether concerted evolution through recombination posed a plausible explanation for the excessive sequence similarity between chicken CYP1A genes. CODOUBLE is an algorithm that detects levels of paralog-paralog conservation at the third positions of codons which exceed expectations under the assumption of neutral independent evolution. Full-length CYP1A coding sequences from chicken, mouse and human were subjected to pairwise comparisons (e.g. chicken

vs. mouse, mouse vs. human, etc.). Paralog pairs from all species under investigation showed significant evidence of gene conversion (p-values < 10^{-6}).

Spatial heterogeneity in phylogenetic signal is commonly used as an indicator of recombinant sequences. Transitions between sub-sections of an alignment that produce mutually exclusive clustering patterns indicate boundaries of regions with different evolutionary histories. SlidingBayes (Paraskevis *et al.*, 2005) was used to determine if sub-sections of CYP1A coding sequences generated disparate phylogenetic results. Specifically, we examined *a posteriori* support for two clustering patterns – paralog pairs clustered together by species, or CYP1A4 and CYP1A5 clustered with mammalian CYP1A1s and CYP1A2s, respectively (i.e., putative orthologs clustered together). Most of the alignment strongly supports a chicken CYP1A cluster separate from fish or mammalian genes (Figure 2a), the same topology obtained from phylogenetic analyses of full-length sequences. Nucleotide positions 530, 720, 940, and 1260 demarcate regions of the alignment with distinct combinations of ortholog-paralog clustering (Figure 2, grey dotted lines).

The region spanning 721-940 bp presents a phylogeny distinctly different from the full-length. Positions 760-930 bp strongly support clustering of chicken CYP1A4 with mammalian CYP1A1s, and CYP1A5 with mammalian CYP1A2s (Figure 2a). Comparatively weaker support for the CYP1A4/CYP1A1 clade is due to strong similarity in this region between CYP1A4/CYP1A1s and fish CYP1A genes. Support for a clade containing mammalian CYP1A1s, chicken CYP1A4, and fish CYP1As mirrors that for the CYP1A5+CYP1A2 cluster (not shown). Overall, phylogenetic signal from 721-940 bp strongly supports the hypothesis that an ancestral CYP1A1-like gene underwent a single duplication prior to the divergence of the avian and mammalian lineages. The consistency of this phylogeny with the synteny analyses

suggests that this is likely to be the "true" evolutionary history, and that the remaining portions of the chicken CYP1A genes have undergone gene conversion. If so, dips in support for a CYP1A4+CYP1A5 clade, such as those seen at 300-360 bp and 1200-1260 bp (Figure 2a), may reflect boundaries of individual gene conversion tracts.

Surprisingly, SlidingBayes data also suggest gene conversion has impacted the 5' end of mammalian CYP1A genes. The first 530 bp demonstrate strong support for species-specific paralog clusters in both mouse and human, suggesting gene conversion in this region in both lineages (Figures 2b and 3b). Subtle differences in the spatial distribution of support suggest the possibility of different gene conversion tracts in mouse and human (Figure 2b). For example, the sharp dip in support for a human CYP1A clade at 350-400 bp may indicate the presence of one conversion tract spanning 1-350bp and a second at 400-530 bp. In contrast, support for the mouse CYP1A clade escalates gradually, then remains relatively constant out to 500 bp.

In mammals, the region between 550 bp and 940 bp is generally consistent with clustering by ortholog (i.e. a CYP1A1 clade and a CYP1A2 clade) (Figure 2b). Again, weak support for the mammalian CYP1A1 clade in the region of 550-720 bp may be attributable to similarity with fish CYP1As. However, SlidingBayes support for clustering of mammalian CYP1A1s with fish CYP1As is not as strong between 550-720 bp as it is in the subsequent ~250 bp (not shown). Support for clustering of chicken CYP1As with mammalian CYP1A2s at 530-720 bp suggests conversion of CYP1A4 by CYP1A5 (i.e. directional transfer from CYP1A5 to CYP1A4) in the first 720 bp of the chicken paralogs.

The 3' third of the alignment continues to provide support for a mammalian CYP1A1 clade, but support for a CYP1A2 clade drops sharply at 940 bp (Figures 2b and 3e-f). Strong support for an alternative clade(s) involving mouse and/or human CYP1A2s is not evident. Thus, this

drop in support could be attributed to either gene conversion or rapid divergence of this region in mammalian CYP1A2s.

Characterization of recombination break-points and conversion tracts

There are several statistical algorithms designed to pinpoint recombination break-points and define gene conversion tracts. Depending on the algorithm, a combined total of 35-55 recombination events were detected in six genes of interest (i.e., chicken, mouse and human CYP1As). These predicted events were sorted on the basis of statistical significance ($p < 0.01$), and filtered to eliminate redundant predictions and predictions of inter-species recombination (Table 1). The resulting list shows a strong 5' bias (8 of 9 conversion tracts fell in the first ~700 bp) and a slight bias toward transfer of material from the CYP1A2/5 locus to the CYP1A1/4 locus. Recombination also appeared to be more common between chicken CYP1As than between the mammalian paralog pairs.

The predicted gene conversion tracts and recombination break-points were generally consistent with SlidingBayes results and tended to coincide with regions of excessive paralog-paralog sequence similarity. For the chicken genes, the predicted conversion tract spanning 994-1108 bp coincides with a region of >90% identity that encompasses the end of exon 4, all of intron 4, and the first half of exon 5. The conservation of intron 4 is in distinct contrast to other introns, which are generally too divergent to be aligned reliably, and provides strong corroboration for the gene conversion prediction, which was based solely on coding sequence. Similarly, predicted gene conversion tracts spanning 1-297 bp and 370-487 bp in CYP1A4 coincide with regions of >95% identity between CYP1A4 and CYP1A5 (Figure 3), and the region between the two tracts corresponds to a dip in phylogenetic support for a CYP1A4-CYP1A5 clade (Figure 2a).

The predicted recombination break-point at 714 bp in CYP1A4 corresponds to a major drop in support for a CYP1A4-CYP1A5 clade (Figure 2a), and to the end of an extended region of >90% sequence identity. However, the boundaries of several shorter stretches of absolute identity coincide with smaller dips in phylogenetic support for a CYP1A4-CYP1A5 clade, suggesting that the region spanning 1-714 bp encompasses several shorter conversion tracts. As previously noted, break-points at 297 bp and 370 bp are strongly supported by SlidingBayes and paralog-paralog similarity levels. The break-point at 487 bp corresponds roughly to the beginning of a region of reduced similarity (528-548 bp), and a slight dip in sequence similarity coincides with the predicted break-point at 127 bp. Thus, the region spanning 1-714 bp may encompass as many as 4 shorter gene conversion tracts.

Predicted gene conversion events in mouse and human CYP1A genes are confined within the first 535 bp of exon 2, in strong accord with SlidingBayes results. The exact locations of break-points are uncertain as alternative predictions in both mouse and human indicate a range of break-points around 37-52 bp and 522-535 bp. In mouse, CYP1A1-CYP1A2 sequence similarity, SlidingBayes, and RDP2 results all support a single conversion tract spanning the region ~50-520 bp in mouse CYP1A1. In contrast, there is strong support for two distinct conversion tracts covering a slightly longer stretch of human CYP1A1. The predicted break-point at 417 bp in human CYP1A1 is supported by a dip in SlidingBayes support for a human CYP1A1-CYP1A2 cluster between 350-400 bp and a region of reduced sequence similarity at 400-420 bp.

Positive selection in central unconverted region

One possible explanation for the lack of gene conversion in the region 721-940 bp is strong selective pressure for paralog-specific functionality conferred by substrate recognition sites 3 or

4, or the catalytically important I helix (Figure 4). Thus, PAML was used to determine whether the region 721-940 bp has been subject to significant positive selection, as indicated by an excessively high rate of non-synonymous (dN) versus synonymous (dS) substitution. Assuming a tree topology reflecting a single CYP1A gene duplication and allowing for independent substitution rates along all branches, five branches manifest dN/dS significantly >1.0 (Figure 5). Maximum likelihood estimates of dN/dS values for all other branches were <0.5 (not shown). This variable rate branch model was a significantly better fit for the data than a homogeneous model ($2\Delta\ln L = 36.99$, p-value < 0.001). Notably, the strongest evidence of positive selection is in the CYP1A2 lineage after the CYP1A1/CYP1A2 duplication event and before the divergence of birds and mammals. However, the distribution of high dN/dS ratios across the tree (Figure 5) suggests that the region 721-940 bp within tetrapod CYP1As has been subject to strong positive selection throughout its history. As expected, there was no significant evidence for positive selection differentiating CYP1A1/4 from CYP1A2/5 when the full-length coding sequences were considered (data not shown).

DISCUSSION

We have demonstrated that chicken and mammalian CYP1A genes lie in an extended region of conserved fine-scale synteny. Conservation of the region spanning CYP11A1 through CSK suggests that this arrangement was present in the most recent common ancestor of mammals and birds. Furthermore, conservation of the downstream CYP1A-CSK-LMAN1L-ULK3 segment in frog and pufferfish suggests that the "composite" gene arrangement observed on human chromosome 15 is likely to have existed in the common ancestor of mammals and birds. The current arrangement of this region at the end of chicken chromosome 10 likely resulted from an intra-chromosomal rearrangement in the avian lineage subsequent to avian-mammalian

divergence. Indeed, segmental rearrangements are a documented characteristic of the broad-scale synteny between chicken chromosome 10 and human chromosome 15 (Crooijmans *et al.*, 2001). However, it seems unlikely that identical tandem inverted gene duplications occurred independently in both avian and mammalian lineages. Overall, the synteny data strongly suggest a single CYP1A gene duplication event in the tetrapod lineage subsequent to the divergence of amphibians. This in turn implies that avian CYP1A4 and CYP1A5 are orthologous to mammalian CYP1A1 and CYP1A2, respectively. The fact that this orthology is not reflected in traditional phylogenetic analyses (e.g. Gilday *et al.*, 1996a) may be attributed to concerted evolution via inter-paralog gene conversion.

Multiple independent methods indicate that gene conversion has played a major role in CYP1A evolution. Phylogenetic methods are generally considered to be the least sensitive means of recombination detection, but may be most appropriate in cases of frequent gene conversion (Drouin *et al.*, 1999; Posada and Crandall, 2001; Posada, 2002). In contrast, many statistical algorithms lose sensitivity when gene conversion is common and sequence divergence is low (Posada and Crandall, 2001; Wiuf *et al.*, 2001; Posada, 2002). Conversely, those same algorithms may be subject to false positives when presented with highly divergent datasets. However, it is widely recognized that recombination detection algorithms tend to underestimate the incidence of gene conversion in real datasets (Posada and Crandall, 2001; Wiuf *et al.*, 2001; Posada, 2002). In the current case, the concordance of SlidingBayes and RDP2 results and the presence of corroborating stretches of paralog-paralog sequence identity provide a high degree of confidence in predicted gene conversion tracts, even those detected by only one of five statistical algorithms. This point is particularly well illustrated by the predicted gene conversion tract 994-1108 bp in chicken; while MaxChi was the only statistical algorithm to predict this event,

SlidingBayes strongly supports paralog clustering in this region and >90% conservation of intronic sequence in this region makes the predicted gene conversion almost unquestionable.

In cases of extensive conversion, it is difficult to distinguish between repetitive conversion and multiple adjacent conversion events; this predicament is particularly evident in the 5' portion of chicken CYP1As. RDP2 detected one long conversion tract spanning 1-714 bp in CYP1A4, as well as three shorter tracts which overlap the long tract and, in some cases, each other. Gene conversion creates extended regions of sequence identity that can provide templates for further homologous recombination, thereby propagating gene conversion activity. It is possible that the region spanning 1-714 bp is an older conversion tract, and that predicted internal break-points and conversion tracts correspond to smaller, more recent events (Figure 4).

Likewise, the prediction of a break-point at 535 bp in both mouse and human CYP1A genes may reflect an ancestral conversion event spanning 1-535 bp, with internal break-points corresponding to subsequent conversion events (Figure 4). The gradual increase between 1 and 240 bp in phylogenetic support for a mouse paralog clade is also in accord with the idea of an older conversion spanning this area. In the case of human CYP1As, two shorter conversion tracts separated by a region of older conversion would explain the reduced phylogenetic support for a paralog clade at 350-400 bp and the predicted break-point at 417 bp. Alternatively, 522-535 bp may represent a recombination hot-spot that has been involved in redundant conversion events in mouse and human CYP1A genes.

A conservative assimilation of the current data suggests at least 8 gene conversion events in the ~300 million years since the divergence of avian and mammalian lineages, or a gene conversion rate of ~$1x10^{-8}$ per locus per generation (Figure 4). The palindromic arrangement of CYP1A paralogs may contribute to both the frequency and the 5' polarity of conversion (i.e. the

frequency of conversion declines 5' to 3'). Inverted repeats longer than 50 bp have been shown to induce double-strand breaks and homologous recombination in yeast (Nasar *et al.*, 2000), and primate MSY palindromes have a history of recurrent arm-to-arm gene conversion at an estimated frequency well in excess of that observed here (Rozen *et al.*, 2003). The extended distance between CYP1A1 and CYP1A2 may account for the apparently lesser frequency and stronger 5' polarity of gene conversion in mammalian species. Analyses of CYP1A sequences from another avian species (unpublished data) are ongoing and should further elucidate the frequency of gene conversion in tetrapod CYP1A genes.

Given the uncertainty surrounding the boundaries of many gene conversion events, it is difficult to establish a precise mean conversion tract length. However, with one exception, predicted conversion tracts were 100-400 bp in length. This is similar to the conversion tract lengths observed in human Y chromosome palindromes (Rozen *et al.*, 2003) and meiotic crossover hot spots (Jeffreys and May, 2004).

On a different note, the longer CYP1A intergenic region in mammals may also account for differences in transcriptional regulation of chicken and mammalian CYP1As. Chicken CYP1A4 and CYP1A5 tend to be more similarly regulated than CYP1A1 and CYP1A2, and chicken CYP1As tend to exhibit aspects of both CYP1A1- and CYP1A2-like regulation (Mahajan and Rifkind, 1999). Expanded intergenic sequence may have allowed for greater elaboration and differentiation in the promoter regions of CYP1A1 and CYP1A2. If anything, these observations lend weight to the idea that the chicken locus is a less derived, direct predecessor of the mammalian locus, intermediate between a single CYP1A and the largely sub-functionalized CYP1A1 and CYP1A2.

From the standpoint of enzyme function, the exact number and precise boundaries of individual conversion events is irrelevant. The converted regions are relatively clear. It is also clear that the region 720-950 bp has remained unaffected by gene conversion. Interestingly, this region encompasses part of substrate recognition site (SRS) 3 and all of SRS 4, as well as the I helix (Figure 4). SRS 6 also falls in a putatively unconverted region at the 3' end of the genes. This suggests that these SRSs may be particularly important in conferring catalytic functions specific to either CYP1A1/4 (e.g. EROD) or CYP1A2/5 (e.g. UROX), and thus, subject to intense selective pressure that would effectively suppress gene conversion. Innan (2003) demonstrated that the effective rate of gene conversion between the human RH antigen genes is reduced in regions under strong functional selection.

Likewise, our PAML analysis provides strong evidence for positive selection acting on the unconverted region of tetrapod CYP1As. Whereas Innan (2003) based claims of positive selection on dN/dS ratios of ~3:1, dN/dS ratios in excess of 20:1 were evident along several branches in the current work. The highest dN/dS ratio occurs in the CYP1A2/5 lineage subsequent to gene duplication. This finding implies a scenario involving gene duplication followed by neo-functionalization. The fact that elevated dN/dS ratios precede the tetrapod gene duplication event suggests a pre-existing pressure to derive novel functions. The gene duplication may have been a fortuitous event which released CYP1A2 to evolve at a faster pace, thereby facilitating ongoing functional diversification. Overall, this model of CYP1A functional evolution is interesting because the tetrapod CYP1A paralogs have generally been viewed as a case of duplication and sub-functionalization. The continuing positive selection in the mammalian CYP1A lineages is also intriguing, as it suggests mammal-specific pressures to

expand the repertoire of CYP1A substrates. It may also explain the more limited extent of gene conversion in mammalian, as compared to chicken, CYP1A genes.

The position of the ancestral CYP1A locus remains a question. Opposing orientations of the pufferfish and frog CYP1A genes pose a conundrum with regard to determining whether CYP1A1 or CYP1A2 resides in the ancestral CYP1A gene locus. Unfortunately, due to misassembly of the CYP1A gene in the current assembly of the zebrafish (*Danio rerio*) genome, no meaningful information can be derived from that species. To date, there are no other fish or amphibian genome assemblies available to aid in the resolution of this question. However, as frog shared a more recent common ancestor with birds and mammals, it is tempting to speculate that the CYP1A1 locus is ancestral. There is no evidence of a second CYP1A gene elsewhere in the *X. tropicalis* genome, and the presence of two CYP1A genes in *X. laevis* (Fujita *et al.*, 1999) is probably an artifact of tetraploidy in that species. Thus, it is unlikely that this interpretation could be complicated by a CYP1A paralog in *X. tropicalis.* The idea of an ancestral CYP1A1-like locus is in accord with comparisons of sequence identity, enzyme function, and transcriptional regulation which show slightly greater similarity between mammalian CYP1A1s and fish CYP1As (Morrison *et al.*, 1995), but this could also be explained by positive selection on CYP1A2/5 soon after the duplication.

This work emphasizes the need to consider non-canonical evolutionary mechanisms when dealing with CYPs. The cytochromes P450 comprise the largest gene super-family to date, with more than 5,000 known members in bacteria and eukaryotes and as many as 400+ genes in a single species (D. Nelson, pers. comm.). This astounding abundance has, of course, arisen primarily through gene duplication. Tandem duplication arrays are common among CYPs, as are documented cases of inter-paralog recombination (Sinnott *et al.*, 1990; Pascoe *et al.*, 1992).

For example, rat CYP2B genes are located in a tandem array on chromosome 1, and interlocus gene conversion is posited to account for 98% nucleotide identity between CYP2B1 and 2B2 (Atchison and Adesnik, 1986). We suggest that gene conversion between tandemly duplicated CYP genes may be more common than previously recognized, and that gene conversion may obfuscate orthology between CYPs. In the case of chicken CYP1A4 and CYP1A5, there is too great a precedent for the use of the current names to warrant a change. However, it is important to recognize the orthology of the chicken and mammalian loci, particularly the sequence orthology that remains in the unconverted regions. Furthermore, given that CYP nomenclature is intended to reflect evolutionary relationships, appropriate efforts should be made to determine true ortholog/paralog relationships prior to naming of future CYPs.

ACKNOWLEDGEMENTS

Funding for this work was provided by the NIH Superfund Basic Research Program at Boston University (5-P42-ES-07381) and by the Woods Hole Oceanographic Institution.

SUPPLEMENTARY MATERIAL

The alignment of CYP1A coding sequences used for all analyses presented herein is provided in multi-sequence FASTA format (vertebrate_CYP1As.fasta).

LITERATURE CITED

Aguileta, G., J. P. Bielawski and Z. Yang (2004). Gene conversion and functional divergence in the beta-globin gene family. *J Mol Evol* 59(2): 177-89.

Atchison, M. and M. Adesnik (1986). Gene conversion in a cytochrome P-450 gene family. *Proc Natl Acad Sci U S A* 83(8): 2300-4.

Balding, D. J., R. A. Nichols and D. M. Hunt (1992). Detecting gene conversion: primate visual pigment genes. *Proc Biol Sci* 249(1326): 275-80.

Crooijmans, R. P., R. J. Dijkhof, T. Veenendaal, J. J. van der Poel, R. D. Nicholls, H. Bovenhuis and M. A. Groenen (2001). The gene orders on human chromosome 15 and chicken chromosome 10 reveal multiple inter- and intrachromosomal rearrangements. *Mol Biol Evol* 18(11): 2102-9.

Drouin, G., F. Prat, M. Ell and G. D. Clarke (1999). Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol* 16(10): 1369-90.

Elskus, A. A., E. Monosson, A. E. McElroy, J. J. Stegeman and D. S. Woltering (1999). Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? *Aquatic Toxicology* 45: 99-113.

Fujita, Y., H. Ohi, N. Murayama, K. Saguchi and S. Higuchi (1999). Molecular cloning and sequence analysis of cDNAs coding for 3-methylcholanthrene-inducible cytochromes P450 in Xenopus laevis liver. *Arch Biochem Biophys* 371(1): 24-28.

Gilday, D., M. Gannon, K. Yutzey, D. Bader and A. B. Rifkind (1996a). Molecular cloning and expression of two novel avian cytochrome P450 1A enzymes induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin. *J Biol Chem* 271(51): 33054-9.

Gilday, D., M. Gannon, K. Yutzey, D. Bader and A. B. Rifkind (1996b). Molecular cloning and expression of two novel avian cytochrome P450 1A enzymes induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin. *J Biol Chem* 271(51): 33054-9.

Gilday, D., K. Yutzey, D. Bader and A. B. Rifkind (1996c). Molecular cloning of two TCDD-induced chicken CYP1A related enzymes distinct from mammalian CYP1A1 and 1A2. *FASEB Journal* 10(3): A283.

Godard, C. A. J., J. V. Goldstone, M. R. Said, R. L. Dickerson, B. R. Woodin and J. J. Stegeman (2005). The New Vertebrate CYP1C Family: Cloning of New Subfamily Members and Phylogenetic Analysis. *Biochem. Biophys. Res. Comm.* 331: 1016-1024.

Gonzalez, F. J. and S. Kimura (2003). Study of P450 function using gene knockout and transgenic mice. *Arch Biochem Biophys* 409(1): 153-8.

Gorman, N., H. S. Walton, J. F. Sinclair and P. R. Sinclair (1998). CYP1A-catalyzed uroporphyrinogen oxidation in hepatic microsomes from non-mammalian vertebrates (chicken and duck embryos, scup and alligator). *Comparative Biochemistry and Physiology* 121C(1-3): 405-412.

Innan, H. (2003). A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc Natl Acad Sci U S A* 100(15): 8793-8.

Jeffreys, A. J. and C. A. May (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 36(2): 151-6.

Jouvin-Marche, E., M. Heller and S. Rudikoff (1986). Gene correction in the evolution of the T cell receptor beta chain. *J Exp Med* 164(6): 2083-8.

Kimura, S., F. J. Gonzalez and D. W. Nebert (1984). The murine Ah locus: comparison of the complete cytochrome P_1-450 and P_3-450 cDNA nucleotide and amino acid sequences. *J. Biol. Chem.* 259: 10705-10713.

Mahajan, S. S. and A. B. Rifkind (1999). Transcriptional activation of avian CYP1A4 and CYP1A5 by 2,3,7, 8-tetrachlorodibenzo-p-dioxin: differences in gene expression and regulation compared to mammalian CYP1A1 and CYP1A2. *Toxicol Appl Pharmacol* 155(1): 96-106.

Mahata, S. C., R. Mitsuo, J.-y. Aoki, H. Kato and T. Itakura (2003). Two forms of cytochrome P450 cDNA from 3-methylcholanthrene-treated European eel *Anguilla anguilla*. *Fish. Sci.* 69(3): 615-24.

Martin, D. P., C. Williamson and D. Posada (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21(2): 260-2.

Michelson, A. M. and S. H. Orkin (1983). Boundaries of gene conversion within the duplicated human alpha-globin genes. Concerted evolution by segmental recombination. *J Biol Chem* 258(24): 15245-54.

Morrison, H. G., M. F. Oleksiak, N. W. Cornell, M. L. Sogin and J. J. Stegeman (1995). Identification of cytochrome P-450 1A (CYP1A) genes from two teleost fish, toadfish (Opsanus tau) and scup (Stenotomus chrysops), and phylogenetic analysis of CYP1A genes. *Biochem J* 308 (Pt 1): 97-104.

Morrison, H. G., E. J. Weil, S. I. Karchner, M. L. Sogin and J. J. Stegeman (1998). Molecular cloning of CYP1A from the estuarine fish *Fundulus heteroclitus* and phylogenetic analysis of CYP1A genes: Update with new sequences. *Comparative Biochemistry and Physiology* 121C(1-3): 231-240.

Nasar, F., C. Jankowski and D. K. Nag (2000). Long palindromic sequences induce double-strand breaks during meiosis in yeast. *Mol Cell Biol* 20(10): 3449-58.

Okino, S. T., L. C. Quattrochi, H. J. Barnes, S. Osanto, K. J. Griffin, E. F. Johnson and R. H. Tukey (1985). Cloning and characterization of cDNAs encoding 2,3,7,8-tetrachlorodibenzo-p-dioxin-inducible rabbit mRNAs for cytochrome P-450 isozymes 4 and 6. *Proc Natl Acad Sci U S A* 82(16): 5310-4.

Paraskevis, D., K. Deforche, P. Lemey, G. Magiorkinis, A. Hatzakis and A. M. Vandamme (2005). SlidingBayes: exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. *Bioinformatics* 21(7): 1274-5.

Pascoe, L., K. M. Curnow, L. Slutsker, J. M. Connell, P. W. Speiser, M. I. New and P. C. White (1992). Glucocorticoid-suppressible hyperaldosteronism results from hybrid genes created by unequal crossovers between CYP11B1 and CYP11B2. *Proc Natl Acad Sci U S A* 89(17): 8327-31.

Posada, D. (2002). Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol* 19(5): 708-17.

Posada, D. and K. A. Crandall (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* 98(24): 13757-62.

Quattrochi, L. C., S. T. Okino, U. R. Pendurthi and R. H. Tukey (1985). Cloning and isolation of human cytochrome P-450 cDNAs homologous to dioxin-inducible rabbit mRNAs encoding P-450 4 and P-450 6. *DNA* 4(5): 395-400.

Rabergh, C. M., N. H. Vrolijk, M. M. Lipsky and T. T. Chen (2000). Differential expression of two CYP1A genes in rainbow trout (*Oncorhynchus mykiss*). *Toxicol Appl Pharmacol* 165(3): 195-205.

Rifkind, A. B., A. Kanetoshi, J. Orlinick, J. H. Capdevila and C. Lee (1994). Purification and biochemical characterization of two major cytochrome P-450 isoforms induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin in chick embryo liver. *J Biol Chem* 269(5): 3387-96.

Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston, R. K. Wilson and D. C. Page (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423(6942): 873-6.

Rudikoff, S., W. M. Fitch and M. Heller (1992). Exon-specific gene correction (conversion) during short evolutionary periods: homogenization in a two-gene family encoding the beta-chain constant region of the T-lymphocyte antigen receptor. *Mol Biol Evol* 9(1): 14-26.

Sinclair, P. R., N. Gorman, I. B. Tsyrlov, U. Fuhr, H. S. Walton and J. F. Sinclair (1998). Uroporphyrinogen oxidation catalyzed by human cytochromes P450. *Drug Metab Dispos* 26(10): 1019-25.

Sinnott, P., S. Collier, C. Costigan, P. A. Dyer, R. Harris and T. Strachan (1990). Genesis by meiotic unequal crossover of a de novo deletion that contributes to steroid 21-hydroxylase deficiency. *Proc Natl Acad Sci U S A* 87(6): 2107-11.

Teraoka, H., W. Dong, Y. Tsujimoto, H. Iwasa, D. Endoh, N. Ueno, J. J. Stegeman, R. E. Peterson and T. Hiraga (2003). Induction of cytochrome P450 1A is required for circulation failure and edema by 2,3,7,8-tetrachlorodibenzo-p-dioxin in zebrafish. *Biochem Biophys Res Commun* 304(2): 223-8.

Teshima, K. M. and H. Innan (2004). The effect of gene conversion on the divergence between duplicated genes. *Genetics* 166(3): 1553-60.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25(24): 4876-82.

Wiuf, C., T. Christensen and J. Hein (2001). A simulation study of the reliability of recombination detection methods. *Mol Biol Evol* 18(10): 1929-39.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13(5): 555-6.

Yawetz, A., B. R. Woodin and J. J. Stegeman (1998). Cytochromes P450 (CYP) in liver of the turtle *Chysemys picta picta* and the induction and partial purification of CYP1A-like proteins. *Biochim. Biophys. Acta.* 1381(1): 12-26.

Table 1. Summary of intraspecies conversion events detected with $p < 0.01$ by any of five algorithms (RDP, GeneConv, MaxChi, Bootscan, or Chimaera) were considered. In cases where multiple methods detected the same or a similar conversion event, we report consensus recombination boundaries, concurring methods, and the best p-value. Bold print denotes the method yielding the best p-value. Two fish CYP1A sequences (*D. rerio* and *F. heteroclitus*) were included as external non-recombinant reference sequences.

SPECIES	CONVERSION DIRECTION	REGION (NT)	BEST P-VALUE	METHODS
G. gallus	CYP1A5 → CYP1A4	1-714	1.58E-13	GeneConv
G. gallus	CYP1A5 → CYP1A4	2-297	3.54E-08	**MaxChi**, GeneConv
G. gallus	CYP1A5 → CYP1A4	370-487	1.36E-07	MaxChi
G. gallus	CYP1A4 → CYP1A5	1-127	1.35E-04	GeneConv
G. gallus	CYP1A4 → CYP1A5	994-1108	2.17E-03	MaxChi
M. musculus	CYP1A2 → CYP1A1	37/46-522/8	1.49E-22	**MaxChi**, RDP, GeneConv, Bootscan, Chimaera
H. sapiens	CYP1A2 → CYP1A1	37-522	5.71E-09	**MaxChi**, RDP, GeneConv, Bootscan
H. sapiens	CYP1A2 → CYP1A1	417-534	6.83E-09	**MaxChi**, GeneConv, Bootscan, Chimaera

Figure 1. Illustration of gene order and orientation in regions surrounding CYP1A gene(s) on human (*Homo sapiens*) chromosome 15, chicken (*Gallus gallus*) chromosome 10, frog (*Xenopus tropicalis*) scaffold 287, and an unspecified pufferfish (*Tetraodon nigridis*) chromosome. Gene lengths and intergenic distances are not drawn to scale. Double slashes (//) indicate continuing sequence data. The physical end of chicken chromosome 10 is indicated by an arrowhead (◄). The end of sequence data for *X. tropicalis* scaffold 287 is indicated by a circle (●); the physical location of this sequence within the frog genome is undetermined.

Figure 2. Spatial distribution of posterior probability support for specific phylogenetic clusters, as determined by SlidingBayes. Grey dotted lines indicate partition boundaries utilized for subsequent analyses. Note: clusters consisting of CYP1A5 + CYP1A1s and CYP1A4 + CYP1A2s received no significant support and, thus, are not shown.

Figure 3. Gene structure and sequence conservation for chicken (top) and human (bottom) CYP1A genes. Coding exons are indicated by solid rectangles and 5' untranslated exons by open rectangles. Chicken introns (solid lines) are shown to scale with alignment gaps delimited by slashes. Human introns (dotted lines) are not drawn to scale. Regions that are unalignable (ua) or for which there is no sequence data (no data, nd) are shown in grey. Percent identity values are shown between the two genes.

Figure 4. Graphical summary of predicted gene conversion tracts and their positions relative to structural and functional domains of CYP1A proteins. Substrate recognition sites (SRSs) are shaded black (top track). α-Helices (cylinders, letters) and β-sheets (arrows, numbers) are shown in the second track. Conversion tracts supported by both SlidingBayes and RDP2 are shown in black. Gene conversion events supported only by RDP2 are shown in grey. Converted regions suggested by SlidingBayes but not statistically supported by RDP2 are open.

Figure 5. Evidence of positive selection in tetrapod CYP1A lineages, as determined by PAML. Branches with dN/dS ratios significantly >1.0 are shown in bold red with dN/dS values below the branch. dN/dS values <1.0 are not shown.

Figure 1

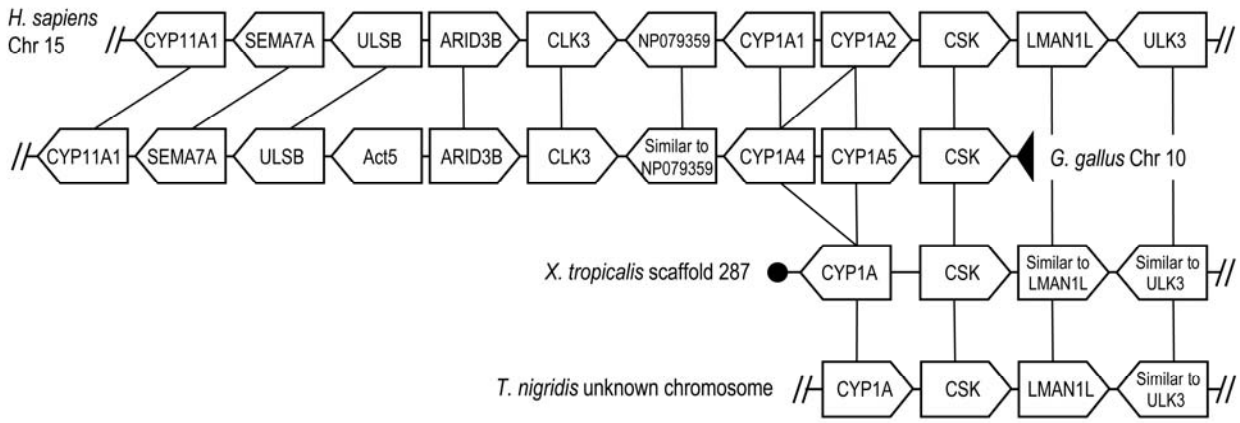

H. sapiens
Chr 15
CYP11A1
SEMA7A
ULSB
ARID3B
CLK3
NP079359
CYP1A1
CYP1A2
CSK
LMAN1L
ULK3
CYP11A1
SEMA7A
ULSB
Act5
ARID3B
CLK3
Similar to NP079359
CYP1A4
CYP1A5
CSK
G. gallus Chr 10
X. tropicalis scaffold 287
CYP1A
CSK
Similar to LMAN1L
Similar to ULK3
T. nigridis unknown chromosome
CYP1A
CSK
LMAN1L
Similar to ULK3

Figure 2

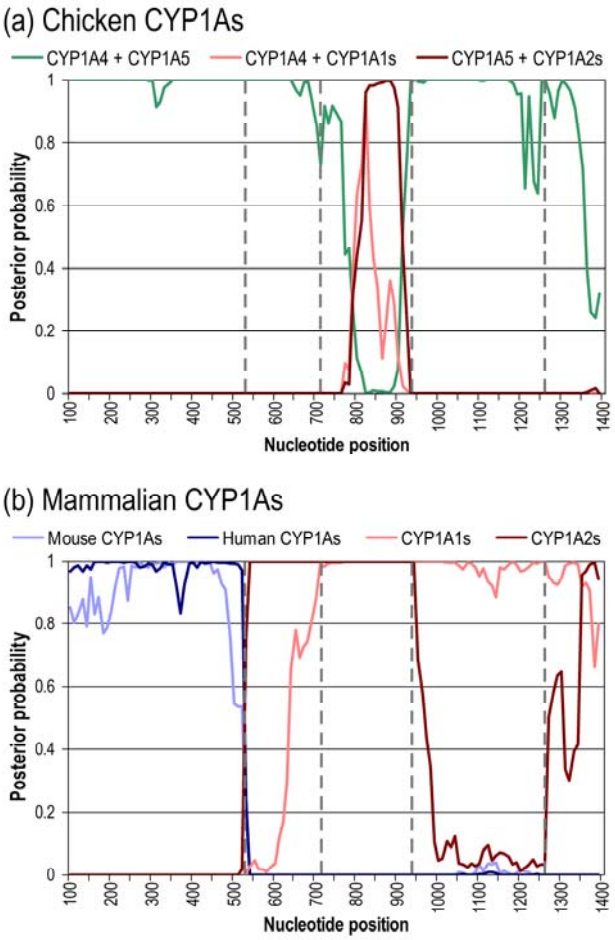

(a) Chicken CYP1As
CYP1A4 + CYP1A5
CYP1A4 + CYP1A1s
CYP1A5 + CYP1A2s
Posterior probability
Nucleotide position
(b) Mammalian CYP1As
Mouse CYP1As
Human CYP1As
CYP1A1s
CYP1A2s
Posterior probability
Nucleotide position

Figure 3

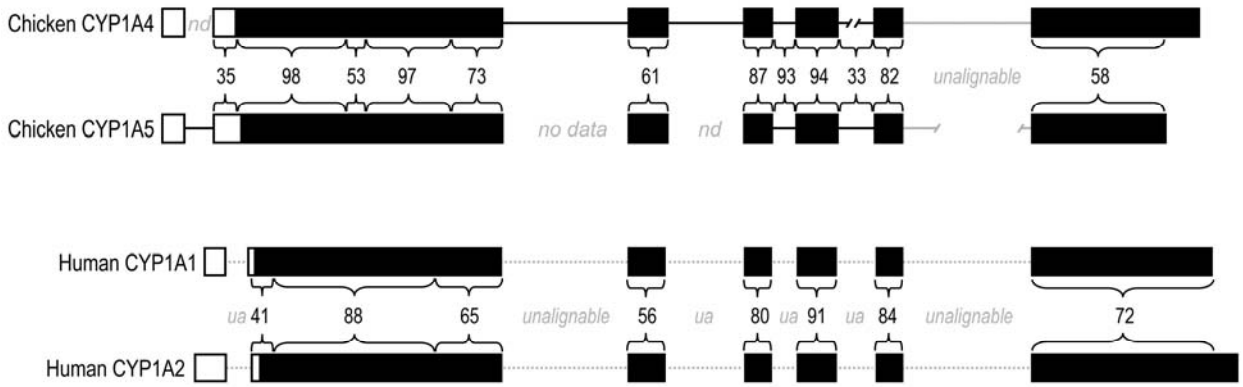

Chicken CYP1A4
nd
35
98
53
97
73
61
87
93
94
33
82
unalignable
58
Chicken CYP1A5
no data
nd
Human CYP1A1
ua
41
88
65
unalignable
56
ua
80
ua
91
ua
84
unalignable
72
Human CYP1A2

Figure 4

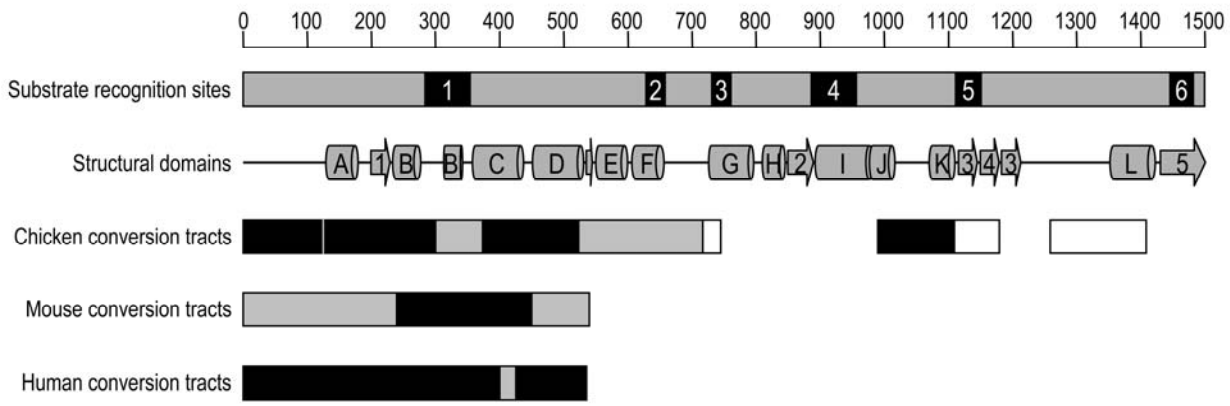
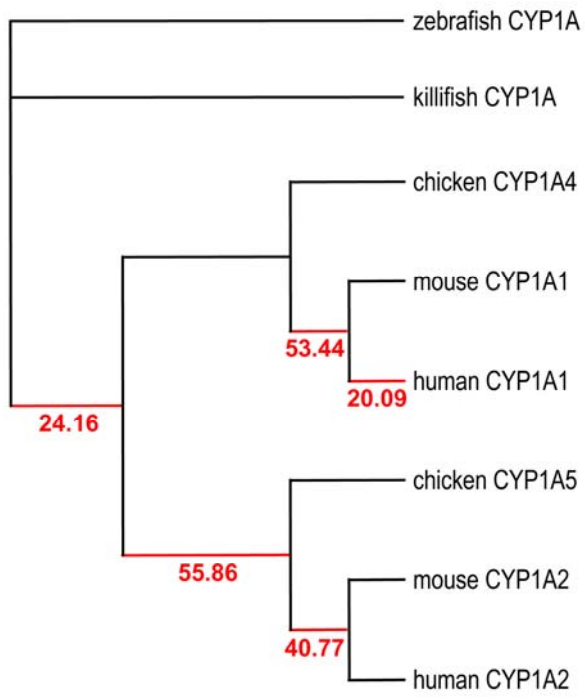

0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
Substrate recognition sites
1
2
3
4
5
6
Structural domains
A
1
B
B
C
D
E
F
G
H
2
I
J
K
3
4
3
L
5
Chicken conversion tracts
Mouse conversion tracts
Human conversion tracts

Figure 5


zebrafish CYP1A
killifish CYP1A
chicken CYP1A4
mouse CYP1A1
53.44
human CYP1A1
20.09
24.16
chicken CYP1A5
55.86
mouse CYP1A2
40.77
human CYP1A2