PHYLOGENETIC AND FUNCTIONAL ANALYSES OF CYP FAMILIES 2 AND 4

# PHYLOGENETIC AND *IN SILICO* FUNCTIONAL ANALYSES OF THE CYTOCHROME P450 FAMILIES 2 AND 4

By

NINA L. KIRISCHIAN, B.Sc.

A thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2011) McMaster University

(Biology) Hamilton, Ontario

TITLE: Phylogenetic and *in silico* Functional Analyses of the Cytochrome P450 Families 2 and 4

AUTHOR: Nina L. Kirischian, B.Sc. (McMaster University)

SUPERVISOR: Dr. J. Y. Wilson

NUMBER OF PAGES: xiii, 196

## ABSTRACT

The CYP superfamily is present in all domains of life and can metabolize an array of exogenous and endogenous compounds. In vertebrates, the CYP2 family is the largest and most diverse family; in mammals this CYP family is important for drug metabolism. The CYP4 family is the major family involved in the metabolism of fatty acids and eicosanoids. Both families have uncertain phylogenetic relationships amongst the vertebrate subfamilies and functional studies in non-mammalian vertebrates are limited. Vertebrate CYPs usually present a one-to-one subfamily relationship, yet in the CYP2 and in some CYP4 subfamilies this does not hold true, thus extrapolating functional understanding is difficult. Phylogenetic trees were constructed for the CYP2 and CYP 4 families using robust phylogenetic methods (maximum likelihood and Bayesian inference) and sequences from all vertebrate classes. An emphasis was placed upon using the full complement of the family from species with completed genomes, particularly from mammalian and actinopterygian species. In the CYP2 phylogeny, subfamilies from distinct vertebrate classes were rarely clustered. CYP substrate recognition sites (SRSs), regions previously proposed as important for determining differences in substrate specificity between CYP genes, did not have elevated rates of evolution in either CYP2 or CYP4 analyses. The CYP4 phylogeny supported the placement of the CYP4V clade with invertebrate sequences and resolved inconsistencies in the placement of mammalian CYP4A/4X/4Z and 4B subfamilies. The evolution of the CYP4T and CYP4B subfamilies will need more consideration. *In silico* functional analyses raised testable hypotheses regarding the CYP4X and CYP4F22 genes. CYP4X

was hypothesized to selectively metabolize long chain fatty acid amides while CYP4F22 was proposed to have overlapping function with other CYP4F genes with metabolic activity towards leukotrienes. Functional testing within the CYP4F clade should begin with long chain fatty acids and eicosanoids, such as leukotriene $B_4$.

## ACKNOWLEDGEMENTS

This thesis would not have been possible if not for the amazing people that surrounded me throughout this journey. I would like to express my since gratitude to my amazingly supportive and encouraging family. Thank you for believing in me and helping me succeed in my goals. A special thank you goes to my supervisor, Dr. Joanna Wilson, for her extensive guidance with my dynamic projects and for her support throughout my Masters experience. Thank you for introducing me to the wonderful world of cytochrome P450 and so much more, the lessons learned will be invaluable for my future academic and life achievements. A very special thank you is owed to Dr. Andrew G. McArthur. Your mentorship and teaching have been the highlight of my first year and your critical guidance throughout my second project was essential for the successful completion. Thank you Andrew and Joanna for teaching me the basics of programming and the exciting aspects of evolutionary biology and microarray chip design. Thank you Dr. Brian Golding, for your continuous advice, "hypothetical questions", and encouragement during the thesis/manuscript revision process. Thank you to the awesome Wilson Lab (family), and to the resourceful Golding Lab for teaching me so much about different aspects of biology and the fun world of Unix and Perl, respectively. In Particular, a very special thank you is owed to Marcus S., Melanie L., Alicia D., Emily S., Sinah L., Luke B., Barb R., and Wilfried H., for their incredible and consistent support, encouragement, editing, and advice which made this journey so much more fun and rewarding. Marcus: you were always the first to help and provide amazing advice; your help will never be forgotten. Melanie: Your friendship, teaching and our

baking Wednesdays will be with me forever and your generous help will not be forgotten, I am so very grateful. Sinah: you were my LSB mom keeping me on track and always encouraging me to push forward. Emily: you were the best Wilson Lab big sister I could have ever asked for, thanks for being so encouraging and positive. Luke: You were by far the best BGSS Co-President I could have ever asked for and equally an amazing big brother of the lab and for that my thanks will always come in éclair form to you. To the rest of the lab, thank you for being so fantastic and always enjoying my baking/cooking experiments, it truly has been a PLEASURE!

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

**Appendix 1: Supplementary Figures and Tables for Chapter 2**



**Appendix 2: Supplementary Figures and Tables for Chapter 3**

## ABBREVIATIONS

$\alpha$ : Gamma shape parameter

**AFBI:** Aflatoxin BI

**AIC:** Akaike Information Criteria

**BIC:** Bayesian Information Criterion

**BLAST:** Blast Local Alignment Search Tool

**Bp:** base pair

**CYP (or P450):** Cytochrome P450

**EST:** Expressed Sequence Tags

**F:** Amino acid frequency

**FAAH:** Fatty acid amide hydrolase

**G:** gamma rate

**I:** Invariant sites

**HETE:** Hydroxyeicosatetraenoic acid

**LRT:** Likelihood ratio test

**MAM:** Mammalian/ Mammals

**MC3:** Metropolis coupled Markov chain Monte Carlo

**ML:** Maximum likelihood

**NCBI:** National Center for Biotechnology Information

**PAUP:** Phylogenetic Analysis Using Parsimony

**PDB:** Protein Data Bank

**RAxML:** Randomized axelerated Maximum Likelihood

**RefSeq:** Reference Sequence

**RNA:** Ribonucleic Acid

**SAS:** Statistical analysis software

**SRS:** Substrate recognition sites

**TPM:** Transcripts per Million

**VLFA:** Very Long Fatty Acids

# CHAPTER 1:

# GENERAL INTRODUCTION

## 1.1 Cytochrome P450 Superfamily

### *1.1.1 Cytochrome P450 Evolution*

The cytochrome P450 (CYP) superfamily is found in all domains of life (Bernhardt 2006; Nebert and Dalton 2006; Nelson 1998; Nelson 2010). The total number of described CYP genes has significantly expanded to over 12,000 sequences, divided into 977 families (Nelson 2009; Robert *et al.* 2010). CYP sequence numbers will continue to expand with new genome annotations of plants, fungi, archaea, protists, and bacteria. Currently plant (i.e. *Zea mays*, tomato, and potato) and sea anemone (i.e. *Nematostella vectensis* and *Lottia gigantea*; Nelson 2010) are targeted for CYP gene identification. CYP proteins can metabolize an array of exogenous (i.e. pharmaceuticals drugs, pollutants and organic compounds) and endogenous (i.e. lipids, eicosanoids, and vitamin D) compounds (Lepesheva and Waterman 2004; Nebert and Dalton 2006; Simpson 1997; Thomas 2007). Of all CYP families, CYP51 is the only CYP family found in all domains of life (Lepesheva and Waterman 2007; Yoshida *et al.* 2000) and is hypothesized as the most ancient CYP family (Aoyama *et al.* 1996; Nelson 1999) with a functional importance of 14α-demethylation of sterols (Lepesheva and Waterman 2007). Phylogenetic analyses suggest placement of the CYP51 family as the basal group of CYP relationships (Nelson 2009; Yoshida *et al.* 2000).

*1.1.2 Nomenclature and Genome Sequencing*

The large diversity of CYP sequences has been placed in a hierarchical structure of clans, families and subfamilies based on sequence similarity and phylogenetic topology. CYP clans organize CYP families into hierarchal deep branching clades according to the common ancestor by using phylogenetic reconstructions (Nelson 1999; Nelson 2004; Nelson 2010). In animals, there are 11 identified clans that include the CYP2, CYP3, CYP4, CYP7, CYP19, CYP20, CYP26, CYP46, CYP51, and CYP74 clans and the mitochondrial clan (Nelson 2010). Some clans are composed of multiple CYP families, whereas others are composed of a single expanded family (Nelson 2003; Nelson *et al.* 2004). For example, clan 2 contains the CYP21, CYP17, CYP1 and CYP2 families whereas, clan 4 contains only the CYP4 family genes (Nelson 2003; Nelson *et al.* 2004). Based on the topology of clans, the evolutionary history of CYPs can be hypothesized and the general CYP evolution in metazoans has been summarized in Nelson *et al.* (1998; 1999). At the family level, there are 18 CYP families in vertebrates that mostly have a one-to-one homology across taxonomic classes; in rare cases CYP families are specific to certain lineages (Nelson 2003). The CYP39 family was found in mammals but has not been identified in fish: the only CYP family that was not shared in both vertebrate classes (Nelson 2003). The only CYP39 function known to date is 7α- hydroxylation of 24-hydroxycholesterol (Li-Hawkins *et al.* 2000). At the subfamily level, vertebrate subfamilies in CYP families 1-4 have significant diversity and subfamilies in these clans may be specific to one taxonomic class (Nelson 2003). These subfamilies have a primary role in exogenous compound metabolism; those CYP families whose primary role is in

production and metabolism of endogenous compounds tend to have shared subfamilies and nearly a one-to-one correspondence of genes across vertebrate species (Nelson 2003; Thomas 2007).

Genome sequencing has resulted in a massive expansion of CYP gene identification, from 3 to 12,456 sequences between 1982 and 2010 (Nelson 2010). The CYP nomenclature system, proposed by Nebert *et al.* (1987), was based on amino acid similarity between sequences. CYP genes are placed into families (>40% amino acid identity, denoted by a number) and subfamilies (>55 amino acid identity, denoted by a letter) and genes are typically numbered in order of their discovery (Nelson 2003). For example, the CYP4F11 gene was the 11$^{th}$ gene identified within the 'F' subfamily of the fourth family. Nomenclature assignment for newly identified CYP genes is decided by a nomenclature committee and these sequences are released for public access on the P450 homepage website (http://drnelson.uthsc.edu /cytochrome P450.html).

In cases when assignment of a novel CYP gene to a family or subfamily is difficult, phylogeny may assist with nomenclature decisions (Nelson *et al.* 2004). The nomenclature rules can prove difficult for the assignment of all genes, thus, factors such as location in the genome have been suggested for consideration (Nelson *et al.* 2004). Largely expanded CYP gene numbers are seen within some subfamilies and they appear clustered in the genome, derived from tandem duplication events, providing additional data for nomenclature assignment (Nelson *et al.* 2004). Synteny

has also been examined to guide nomenclature but there are difficulties with broad application of synteny for nomenclature (Nelson 2010). Recent analysis of the *C.elegans* and the *Drosophila* genomes found high rates of synteny loss and a scrambled appearance of the genomes (Nelson 2010), limiting the use of synteny for nomenclature in at least these species. However, synteny and clustering was found in the mouse and human genome for CYP2A, CYP2B and CYP2F subfamilies (Hoffman *et al.* 1995; Nelson *et al.* 2004). In general, the phylogenetic reconstructions support the nomenclature system in most vertebrate CYP families (Goldstone *et al.* 2010; Nelson 2003; Nelson 2009; Nelson *et al.* 2004).

### *1.1.3 P450 protein structure*

A CYP coding gene is composed of 1500 bp on average and translates into a 500 amino acid protein (Nelson *et al.* 2004). CYPs are membrane-associated proteins located in the inner membrane of mitochondria or are bound to the endoplasmic reticulum (Williams *et al.* 2000). Active CYP enzymes within a complex with bound CO produce a peak at 450nm in the UV spectrum (Garfinkel 1958; Omura and Sato 1964), this unique absorption provided the name for this superfamily of proteins (Omura and Sato 1964). The first crystal structure of a CYP protein, P450cam CYP101, identified by Poulos *et al.* (1985), included multiple helices (A-L) and beta sheet domains that form a common tertiary structure that is shared amongst the CYP proteins (Graham and Peterson 1999; Hasemann *et al.* 1995; Li and Poulos 1997). Figure 1 shows a homology model of the CYP2D6 protein with the tertiary organization. CYP proteins have a strongly conserved region surrounding the heme core structure and possess poorly conserved N- and C-

termini regions (Hasemann *et al.* 1995). The more conserved regions included helices D, E, K, L, J and the C-terminus of helix I (Graham and Peterson 1999; Hasemann *et al.* 1995; Li and Poulos 1996). The highly variable regions of CYP proteins are associated with the active site and areas important for substrate specificity and these include helices A, B', F, G, H and N-terminus of helix I (Graham and Peterson 1999; Hasemann *et al.* 1995; Li and Poulos 1996). Crystal structures have been determined for several CYP genes including the bacterial P450cam, P450terp, P450 $_{BM-3}$ (Hasemann *et al.* 1995) proteins, rabbit CYP2C5 (Williams *et al.* 2000), and human CYP2D6 (Rowland *et al.* 2006). The overall crystal structure was similar throughout these reconstructions even though there is less than 15% amino acid sequence identity (Hasemann *et al.* 1995; Nelson *et al.* 1993).

### *1.1.4 Substrate Recognition Sites*

Residues that interact with the substrate in the *Pseudomonas putida* P450 101A protein were identified with X-ray crystallography (Poulos *et al.* 1987). Gotoh (1992) aligned 51 sequences from the CYP2 family as well CYP1 and CYP3 sequences and 8 bacterial sequences, including *P. putida* P450 101A, and identified regions with high non-synonymous amino acid substitutions that correlated with the substrate binding sites in P450 101A (Gotoh 1992). Six substrate recognition sites (SRSs) were proposed based on the co-occurrence of high rates of amino acid substitution, residues important for substrate binding in P450 101A, and location of functionally important point mutations in mammalian CYPs (Gotoh 1992). The locations of the SRS regions within the protein are shown in Figure 1. Several of the SRSs overlap with the highly variable regions in the

CYP proteins. The SRS1 consist of the entire B' helix, the B-B'(C-terminus) and the B'-C (N-terminus) loops. The substrate access channel is associated with helices F, G and the F-G loop which include SRS2 (C-terminus of helix F) and SRS3 (N-terminus of G; Gotoh 1992; Graham and Peterson 1999). SRS5 is primarily associated with β1-4 sheets, which are found between the K-K' helices and SRS6 is found in the β4-1 and β4-2 loop. Interestingly, SRS4 is associated with the more structurally conserved helix I; however the N-terminus to the center of the helix, where SRS4 is mapped, has more variability (Gotoh 1992; Graham and Peterson 1999). These SRS regions have been proposed as critical for defining substrate specificity for individual CYP isoforms (Gotoh 1992).

Regions with high rates of codon or amino acid substitution can be identified via *in silico* analyses within a phylogenetic context, by determining evolutionary rates of functional divergence (Gu 2006). Recent studies of evolutionary rates of functional divergence in the mammalian CYP2C subfamily (da Fonseca *et al.* 2007), vertebrate CYP3A (McArthur *et al.* 2003) and CYP1 family in early deuterostomes and vertebrates (Goldstone *et al.* 2007) correlated highly divergent regions of the CYP proteins with some, but not all, SRS regions. Even when the SRS regions were correlated with areas with high evolutionary rates of functional divergence, the boundaries of the regions were not the same (McArthur *et al.* 2003). The inconsistency between the identified SRS regions and regions with high evolutionary rates of functional divergence determined using DIVERGE (Gu and Vander Velden 2002) suggest that the substrate binding active sites are likely not restricted to the regions identified to have high non-synonymous substitution of codons.

*1.1.5 CYP2 family*

The CYP2 family has significantly diversified with 29 subfamilies making this the least conserved family in vertebrates (Nelson 1998; Nelson 2003). CYP2 genes have been identified in *Ciona* and sea urchin species (Goldstone *et al.* 2007), suggesting that this family predates vertebrate evolution. The large diversity of the CYP2 proteins may be at least partially explained by their major role in the metabolism of foreign compounds (e.g. drugs) and more limited role in production or metabolism of endogenous compounds (e.g. steroid metabolism; Nelson 2003). Two subfamilies, CYP2U and CYP2R, are found in all vertebrate classes; the majority of other CYP2 subfamilies are specific to a single vertebrate class (Nelson 2003; Nelson 2009). A previous study by Nelson (2003) determined the evolutionary relationship between human and fugu CYP genes and suggested that some CYP2 actinopterygian and mammalian subfamilies had a shared evolutionary history. The relationships within the CYP2 family were not clear for most of the subfamilies, excluding the CYP2U and CYP2R subfamilies, which were basal to the rest of the subfamilies (Nelson 2003; Thomas 2007).

Outside of the mammalian class, the knowledge of gene function and evolutionary history of the CYP2s is limited. A few studies have begun to identify function of actinopterygian CYP2 genes, but large gaps still exist. The actinopterygian CYP2P and CYP2N subfamilies likely shared a common ancestor gene and function with mammalian CYP2J (Nelson 2003; Oleksiak *et al.* 2003). The metabolic activity of CYP2P (Oleksiak *et al.* 2003) and CYP2N (Oleksiak *et al.* 2000) have been compared to CYP2J

mammalian genes and are suggested to play a role in fatty acid metabolism. Killifish (*Fundulus heteroclitus*) CYP2P and CYP2N genes were primarily expressed in liver and intestine and they could oxidized arachidonic acid (Oleksiak *et al.* 2003). Gene expression of CYP2P genes were decreased because of starvation (Oleksiak *et al.* 2003), but it did not affect CYP2N expression (Oleksiak *et al.* 2000). The sites of expression, function, and altered expression of CYP2P genes were similar to those of mammalian CYP2J genes (Oleksiak *et al.* 2003). In the actinopterygian CYP2K subfamily, function has been assessed and was not conserved across species. For example, zebrafish CYP2K6 metabolized aflatoxin BI (AFBI), but not lauric acid; yet rainbow trout CYP2K1 metabolized lauric acid but not AFBI (Wang-Buhler *et al.* 2005). This illustrates that functional differentiation can occur in vertebrate class specific subfamilies and yet similar function could be detected in evolutionary separated CYP2 subfamilies. Overall, functional understanding of most non-mammalian subfamilies remains unknown and requires significant study. Since functional hypotheses are often based on assuming functional conservation in orthologs; an understanding of the evolutionary relationships within protein families can provide important information for beginning functional characterization of genes. Considering the significant functional data available for mammalian CYP genes, the completion of a detailed CYP2 family phylogeny for vertebrate sequences may raise functional hypotheses for CYP2 genes in non-mammalian species.

### *1.1.6 CYP4 family*

The CYP4 family is believed to have evolved approximately 1.25 billion years ago, following the development of steroid biosynthesis genes (Simpson 1997). Supporting this hypothesis, three bacterial genes were found to cluster with the CYP4 genes in clan 4 (Nelson 1998) and one of these bacterial genes, *Bacillus megaterium* CYP102, had similar function to eukaryotic CYP4s (Denison and Whitlock 1995). CYP4 protein function has been associated strongly with endogenous compound metabolism; metabolism of exogenous compounds is more limited. CYP4 enzymes are known to primarily ω-hydroxylate the terminal carbon of fatty acid chains to prevent lipotoxicity, and mediate leukotriene and prostanoid catabolism (Hardwick 2008; Hsu *et al.* 2007). Collectively, this suggests that fatty acid hydroxylase activity was one of the most primitive CYP functions (Nelson 1998).

The CYP4 family is divided into 72 subfamilies with more than two thirds, 50 subfamilies, identified in insects and 7 subfamilies in vertebrates (Nelson 2009). In insects, the CYP4 family expanded in the number of subfamilies and this expansion is a strong indicator of functional diversity and adaptation to their diet (Baldwin *et al.* 2009; Davies *et al.* 2006; Feyereisen 2006). Several phylogenies have been completed with vertebrate CYP4 sequences (Fujita *et al.* 2004; Nelson 2003; Thomas 2007). Overall, the topology placement of CYP4V and CYP4F subfamilies were in agreement (Fujita *et al.* 2004; Nelson *et al.* 2004; Thomas 2007). The CYP4V subfamily, one of the few subfamilies shared between vertebrate and invertebrate species, clustered in close proximity with invertebrate CYP4 sequences rather than vertebrate CYP4 subfamilies (Fujita *et al.* 2004; Nelson 2009).

The evolutionary relationships of the majority of CYP4 subfamilies in vertebrate species and of invertebrate to vertebrate subfamilies are unclear. While in most studies the CYP4A, CYP4X, CYP4Z and CYP4B subfamilies cluster, the internal topology has differed (Fujita *et al.* 2004; Nelson 2003; Nelson *et al.* 2004; Thomas 2007), however agreement was found with the actinopterygian CYP4T clade being placed basal in the cluster (Nelson 2003; Thomas 2007). The placement of CYP4T amphibian sequences is unclear; these sequences have resulted in a trichotomy cluster with CYP4T actinopterygian and CYP4B mammalian (Liu *et al.* 2009), or clustered with CYP4B mammalian and avian species (Thomas 2007).

The CYP4 enzymes in vertebrate species catalyze the $\omega$-hydroxylation of the terminal carbon of fatty acids, including essential signaling molecules such as eicosanoids, prostaglandins and leukotrienes (Hardwick 2008). Three subfamilies, CYP4A, CYP4B and CYP4F, have been studied in great detail with regards to their gene expression and function (Hardwick 2008; Simpson 1997); other subfamilies function and metabolic capabilities are less clear. CYP4 studies have been primarily focused on mammalian species and functional interpretation has been limited to mammals, these interpretations are used to infer function for homologous genes in other vertebrate classes. Gene expressions of homologs within CYP4 vertebrates have shown strong variability in their expression, suggesting functional differences between the species. Differences are even found in subfamilies that are restricted to one vertebrate class; CYP4X has variability in both metabolic capabilities and gene expression patterns

between human and mouse (Al-Anizy *et al.* 2006; Savas *et al.* 2005). Similar to CYP2s, the functional and phylogenetic understanding of the CYP4 subfamilies is unclear.

## 1.2 Methodological Approaches for Protein Evolution

### *1.2.1 Phylogenetic methods*

Phylogenetic trees are based on mathematical models that attempt to reconstruct the evolutionary history of specific traits/genes. When selecting the most appropriate phylogenetic method, key factors must be considered such as efficiency, consistency, statistical power, and robustness in the phylogenetic reconstruction (Barton *et al.* 2007; Eisen 1998). Special consideration must be made towards the type of data utilized for the molecular phylogeny (DNA or protein), since it will affect the choice of which substitution model is utilized. All of these factors should be strategically considered before proceeding with the reconstruction.

Selection of substitution models is essential; unlike DNA substitution models, most amino acid substitution models are empirically derived. The empirically derived replacement rates for amino acids (Abascal *et al.* 2005) have been modified from the primary predictions by Dayhoff *et al.* (1978). The JTT substitution model was based on an average of many different protein families with 85% sequence similarity (Jones *et al.* 1992), while the WAG substitution model was based on a maximum likelihood estimation model of 182 protein families that are distantly related (Whelan and Goldman 2001).

Selection of the most appropriate protein amino acid replacement model based on a given protein alignment can be determined using ProtTest, which includes an extensive

suite of amino acid substitution models and the results are ranked based on likelihood (Abascal *et al.* 2005). The models log likelihood are statistically tested for their goodness of fit between the models, and likelihood ratio test and performance based approaches can also be used. Other alternatives, such as Akaike information criteria (AIC) and Bayesian Information Criterion (BIC), assist in determining the best model. For example, lower AIC values suggests a better model fit. Most substitution models incorporate evolutionary rate heterogeneity across sites, via estimation of a gamma distribution with gamma shape parameter ($\alpha$) (Yang and Rannala 1997). The evolutionary rates are not constant across sites, positions that evolve at a slower rate are known to be conserved whereas, sites that change at a faster rate are referred to as variable (Nielsen 2005). The variation in evolutionary rates is often interpreted as a consequence of differences in selective constraints (Gaucher *et al.* 2002; Whelan and Goldman 2001), which are associated with multiple factors such as the three dimensional protein structure and the presence of functional domains both at the RNA (splice sites) and protein levels. Inferring evolutionary data should be done with an acknowledgement that a portion of amino acids are invariable (+I; Reeves 1992), or that sites can have an assigned probability to which rate category they belong (+G; Yang 1993), and that amino acid frequencies can vary (+F; Cao *et al.* 1994). These factors are considered before phylogenetic reconstructions are initiated via Bayesian inference or maximum likelihood (ML) to generate the most plausible reconstruction (Barton *et al.* 2007).

The maximum likelihood approach is used to derive the phylogenetic topology and branch lengths that have the highest likelihood of reconstructing a data set given a

substitution model (Felsenstein 2004; Pevsner 2009). The ML is determined for each position in the alignment and the likelihood of each site is determined as the probability of the variation at each site (Baxevanis and Ouellette 2001). In the case of sites with high likelihood they likely have fewer changes between the branches and generate a higher probability for the internal node. Maximum likelihood works well with large datasets, and offers a statistical model that statistically assesses the evolutionary changes among branches. ML is computationally demanding thus instead of conducting a complete search, a heuristic approach is suggested (Pevsner 2009; Stamatakis *et al.* 2005). The strength and stability of the taxon topology of a maximum likelihood tree may be assessed with the bootstrapping method, introduced by Felsenstein et al (1983). During bootstrapping, the characters in the alignment are sampled at random, with replacement, to generate complete new datasets, which are used to construct the phylogenies (Felsenstein 2004); 100 samples are typical (Berry and Gascuel 1996; Hedges 1992). For each new phylogeny, the bootstrapping trees are scored and compared to result in the optimal phylogenetic tree. For each node on the optimal tree, a bootstrapping value identifies how often those taxa were placed together out of all the bootstrapping replicates. The advantageous aspect of ML is the incorporation of substitution models and likelihood framework for determining historical substitution patterns. Several programs such as Randomized Axelerated Maximum Likelihood (RAxML) integrate a few particular heuristics to reduce searching duration (Stamatakis *et al.* 2005). For example, RAxML generates an initial tree using stepwise addition via parsimony and for branch swapping it utilizes Lazy Subtree Rearrangements (LSR; Stamatakis *et al.* 2005);

using a few other heuristics RAxML generates a quick likelihood tree, yet it is not necessarily the same as an optimal tree (Stamatakis *et al.* 2005).

Bayesian inference of a phylogenetic reconstruction strength is measured using posterior probabilities (Huelsenbeck and Ronquist 2001). Via Bayes's theorem posterior probability distribution for the trees is generated by incorporating the likelihood of the tree topology and starting with a flat prior which will use the default prior probability density, known as the flat Dirichlet (Huelsenbeck and Ronquist 2001; Ronquist *et al.* 2005). The posterior probabilities are calculated using a numerical method, Metropolis coupled Markov chain Monte Carlo (MC3; Huelsenbeck *et al.* 2001); these chains sample the tree space in which the posterior probability is approximated for a given sample (Li *et al.* 2000; Mau *et al.* 1999; Yang and Rannala 1997). There are 4 chains of which 1 is cold and the other 3 are incrementally heated; the purpose of the heated chains is to promote good mixing and flattening the distribution to identify isolated peaks easily and allow for the cold chain to travel in-between these peaks faster and avoid getting caught in one peak (Huelsenbeck and Ronquist 2001). Of the 4 chains only the cold chain values are recorded for the sampled generations, the chains can also exchange states. Chains exchange states if the powers of likelihood ratios are in favor of the change (Huelsenbeck *et al.* 2001). Because the chains begin sampling at a random position, on average it's suggested to remove approximately 25% of the sampled frequencies from the cold chain from the analysis which are referred to as burn-in (Barton *et al.* 2007), yet this may present bias for the remaining generations. The burn-in number of sampled frequencies can be determined by identifying the end of where the plateau is reached for

likelihood ratios and one such program that assesses for this is Are We There Yet server (Nylander *et al.* 2008). One of the challenges with Bayesian inference is that there is no defined number of generations to generate the optimal tree, thus the numbers of generations that are considered enough are assumed (Huelsenbeck *et al.* 2001) and then used to generate a consensus tree. Thus, Bayesian Markov chain can run for an infinite amount of time, although if failed during run time the entire process must start from the beginning (Huelsenbeck and Ronquist 2001). Usually, more than one run is suggested for an analysis, two runs are typical to ensure that there is convergence between the runs. The Markov chains search the given state space for the a likely phylogenetic reconstructions, base on the number of generations excluding the burn-in, a consensus tree is generated (Huelsenbeck and Ronquist 2001). The Bayesian inference allows for calculations of posterior probabilities of any group of interest, based on the trees left after burn-in (Barton *et al.* 2007).

In general, posterior probabilities and bootstrapping values from Bayesian inference and maximum likelihood analysis, respectively, are of great importance for interpretation of the tree as they provide information on the level of certainty for the topology, yet they are not interchangeable values and cannot be directly compared (Douady *et al.* 2003). A distinctly positive correlation is associated with the bootstrapping values and posterior probabilities, but the strengths are variable and posterior probabilities are found to be consistently higher than bootstrapping (Douady *et al.* 2003; Huelsenbeck *et al.* 2002; Whittingham *et al.* 2002; Wilcox *et al.* 2002). Nodes with posterior probabilities of $\geq 0.95$ were found to have bootstrapping values of $>70\%$

(Whittingham *et al.* 2002); similarly bootstrap values of >70% were associates with values of high 0.70 to low 0.90 posterior probabilities, thus Bayesian values of greater than 0.80 are considered as strong support (Whittingham *et al.* 2002). Due to the variability in the sampling of methods, it has been noted that maximum likelihood tends to underestimate the confidence in phylogenetic reconstructions (Huelsenbeck *et al.* 2002).

*1.2.2 In silico functional analysis*

Molecular evolutionary approaches can help raise functional hypotheses in homologous genes. DIVERGE 2.0 allows for the determination of functional divergence between pairwise comparisons of protein groups, centered on the shifted evolutionary rates that occur after gene duplication or speciation (Gaucher *et al.* 2002). These evolutionary rates are site-specific and are essential to determining regions with functional divergence, since regions with high amino acid divergence are suggested to be associated with active sites of an enzyme.

For analyses to be conducted with DIVERGE the following input files are required: a protein alignment (FASTA or CLUSTAL format), a phylogenetic tree file (PHYLIP format) generated from the protein alignment, and a protein structure file (PDB format). The protein with the defined crystal structure is aligned via pairwise comparison to the multiple sequence alignment, for the purpose of visualizing residues onto a 3D structure. Consideration must be given to the protein structure file used to generate the best alignment; using a closely related protein or one from the same family would be ideal. DIVERGE requires the identification of 2 or more groups containing a minimum

of 4 sequences; the groups are typically identified as monophyletic clusters of sequences from the phylogenetic tree. Groups are compared, in a pairwise fashion, using the sequences in the multiple sequence alignment and structure alignment for a statistical analysis of functional divergence between the groups (Gu and Vander Velden 2002).

When the statistical analyses are complete for the pairwise comparisons, a statistical summary is provided for each comparison, which includes theta ML (coefficient of functional divergence), standard error of theta, and likelihood ratio test of theta. In addition, site-specific profiles are derived from posterior analysis outputs that illustrate significant amino acid residues that drive the functional divergence within the pairwise comparison. DIVERGE allows for the visualization of the site-specific profile(s) on the multiple sequence alignment or on a protein structure; although the 3D visualization option is quite poor within the software. Thus, regions with high functional divergence can be identified across the primary amino acid strand and in the tertiary folded state.

In type I functional divergence analyses, the coefficients of evolutionary rates of functional divergence (theta, $\theta$) are calculated for each site with the gamma shape parameter for rate variation among sites. The gamma distribution examines evolutionary rates at sites within each cluster of protein or nucleotide sequences, and then is subtracted from another to determine the theta value for each site-specific comparison. Sites with higher rates of substitution are suggested to account for functional differences between the two groups of sequences (Gaucher *et al.* 2002). For example, determined

evolutionary rates within a protein family or superfamily can be used to derive selection patterns, genetic drift and regions of possible functional importance. The interpretation of the results for type I analyses must consider both the cutoff value of theta (equal to or greater than 0.5) and the likelihood ratio test (LRT) of theta, using the chi-square test. In the case when theta is below the cutoff ($\leq 0.5$) yet the LRT of theta is statistically significant ($p<0.05$), the pairwise comparison will be considered statistically significant.

Type II functional divergence determines radical biochemical changes (hydrophobic vs. hydrophilic) across genes within a phylogenetic context (Gu and Vander Velden 2002; Zheng *et al.* 2007). High positive changes indicate specific residues with consistent and radical biochemical changes, suggesting a functional change in the protein at the particular site that may affect protein folding or substrate binding specificity. Visualizing the biochemical changes specifically mapped on a three dimensional protein structure allows for a better spatial comprehension of the tertiary structure folding. For the type II analysis the 20 amino acids were divided upon 4 groups; positively charged (K, R, H), negatively charged (D,E), hydrophobic (A, I, L, M, F, W, V, Y), and hydrophilic (S, T, N, Q, C, G, P; Gu 2006). The designed algorithm included the frequency with which amino acids substitution occurs; radical changes were considered for amino acid changes from one group to another, otherwise they were considered conserved (Gu 2006). Thus, in the case of type II functional analyses, theta values significantly greater than zero are considered as radical biochemical changes.

Both type I and type II functional divergence are *in silico* analyses that provide focus for deriving testable hypothesis with regards to gene group and protein function.

Functional divergence analyses have been completed on several CYP families including CYP3 (McArthur *et al.* 2003; Yan and Cai 2010), CYP19 (Wilson *et al.* 2005), and CYP1 in early deuterostomes (Goldstone *et al.* 2007). Regions of significant evolutionary rates of functional divergence were detected in the CYP3 family and these homologous genes are known to have functional differences (McArthur *et al.* 2003). The detected functional divergence regions did overlap with some CYP3 SRS regions, however regions outside of SRS were also identified (McArthur *et al.* 2003). Whereas, CYP19 genes function as the catalyst for estradiol synthesis; between the two CYP19 genes functional differences amongst vertebrate is very limited and thus no significant type I functional divergence was detected across the protein, indicating no overlap with detected SRS regions. Collectively, these studies suggest that divergence patterns are likely a good reflection of functional diversity of these families.

*1.2.3 Gene expression profile*

Expressed sequence tags (ESTs) are single-pass reads of 200-800 base pairs of cDNA and provide information on the genes expressed in a given tissue (Nagaraj *et al.* 2007; Parkinson and Blaxter 2009). ESTs have become extremely useful in gene identification and annotation for genome projects. Large numbers of EST libraries are often available for species with completed genomes and these data may be used to examine tissue specific gene expression patterns for related genes to aid in deriving functional hypotheses. EST data is widely available in the UniGene database (Pontius *et al.* 2003; Wang and Liang 2003) for many species. With the completion of genomes for at least one species from most major vertebrate classes (Mammalia, Aves, Amphibia, and

Actinopterygii), this approach may raise hypotheses regarding conserved and divergent function for genes found across vertebrate taxa.

In brief, non-redundant EST tissue libraries can be searched for genes of interest and tissue specific expression patterns that may raise functional hypotheses. In the specific case of CYP genes, those genes involved in exogenous compound metabolism are typically expressed in liver, epithelium cells, and in certain extra-hepatic tissues important for absorption and excretion such as kidney, lung and intestines (Al Omari and Murry 2007; Ding and Kaminsky 2003; Mitschke *et al.* 2008). CYP genes involved in endogenous function may also be expressed in these organs but high basal expression may be noted in extra-hepatic organs that are not involved in absorption or excretion processes. For example, while some CYP4 genes are expressed in liver and kidney, they have been detected in the brain, trachea, mammary glands (Hardwick 2008; Hsu *et al.* 2007; Savas *et al.* 2005). CYP4s have important endogenous functions related to fatty acid metabolism (Hardwick 2008; Simpson 1997).

The digital northern data is calculated by utilizing the total number of transcripts for the gene of interest divided by the total number of transcripts in the library, and the normalized values are shown as transcripts per million (TPM; Pontius *et al.* 2003; Wang and Liang 2003). Digital gene expression data has been used to analyze gene expression of novel CYP1 genes in early deuterostomes (Goldstone *et al.* 2007) and of two different fatty acid amide hydrolase (FAAH) expression in placental mammals (Wei *et al.* 2006). Goldstone *et al.* (2007) used an EST library to infer gene expression of novel genes,

CYP1E1 and several genes from the CYP4F subfamily, in *Ciona intestinalis,* to complement gene expression analysis conducted through RT-PCR throughout different time points of life stage development. EST data for CYP1E1 showed direct expression in gonads, digestive glands and in blood cells, as well as expression data for CYP1F3, which was not analyzed by PCR, was found to be strongly expressed in juveniles (Goldstone *et al.* 2007). Overall, EST data added depth in developing an understanding regarding the functional importance and expression of CYP1s outside of vertebrate species by presenting novel expression data.

### 1.3 Research Topics & Hypotheses

The focus of this thesis is to develop a better understanding of the evolutionary relationships and raise functional hypotheses regarding non-mammalian genes in two CYP families, CYP2 and CYP4. The objectives have been organized into two studies, written as papers in chapter 2 and 3.

Chapter 2 contains a study of the vertebrate CYP2 family evolution and *in silico* functional divergence of CYP2 subfamilies. Sequences representing 24 CYP2 subfamilies were utilized for phylogenetic reconstruction using maximum likelihood and Bayesian inference methods. The mammalian CYP2s are known to metabolize an array of exogenous compounds (i.e. drugs); however, little is known of the metabolic functions outside of this class. The focus of this project was to determine the phylogenetic relationships between vertebrate CYP2 subfamilies with an emphasis on mammalian and actinopterygian CYP2 genes. Functional divergence analyses (type I and type II

functional divergence) were conducted to determine whether highly variable regions in the CYP2 subfamilies correlated with substrate recognition sites (SRS; Gotoh 1992) and represented likely functionally important regions of the protein. In general, the evolutionary relationship and *in silico* functional analyses should lead to testable hypotheses for future CYP2 studies. The publication details are as follows; Phylogenetic and Functional Analysis of the Vertebrate Cytochrome P450 2 Family in the issue of Journal of Molecular Evolution: Volume 72, Issue 1 (2011), Page 56.

Chapter 3 investigates the CYP4 family evolutionary relationships within vertebrate subfamilies and between vertebrate and invertebrate genes. This study addresses the inconsistencies in existing topologies of CYP4 reconstructions and the placement of the actinopterygian and amphibian CYP4T genes to aid in nomenclature issues. Invertebrate CYP4 sequences from four classes (Echinoidea, Ascidiace, Insecta, and Gastropoda) were included to investigate the relationship between CYP4 sequences in vertebrate and invertebrate species. Since, the majority of functional understanding is deduced from mammalian data, we conducted *in silico* functional divergence (type I and type II) analysis to indentify similarities and differences of CYP4 subfamilies across taxonomic classes. Lastly, to deduce specific gene expression patterns, ESTs were utilized to explore CYP4 gene expression in four vertebrate classes.

## 1.4 References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-5

Al-Anizy M, Horley NJ, Kuo CW, Gillett LC, Laughton CA, Kendall D, Barrett DA, Parker T, Bell DR (2006) Cytochrome P450 Cyp4x1 is a major P450 protein in mouse brain. Febs J 273:936-47

Al Omari A, Murry DJ (2007) harmacogenetics of the cytochrome P450 enzyme system: review of current knowledge and clinical significance. J. Pharm. Pract. 20:206-218

Aoyama Y, Noshiro M, Gotoh O, Imaoka S, Funae Y, Kurosawa N, Horiuchi T, Yoshida Y (1996) Sterol 14-demethylase P450 (P45014DM*) is one of the most ancient and conserved P450 species. J Biochem 119:926-33

Baldwin WS, Marko PB, Nelson DR (2009) The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. BMC Genomics 10:169

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH (2007) Evolution. Cold Spring Harbour Laboratory Press, Woodbury, New York

Baxevanis AD, Ouellette BFF (2001) Bioinformatics: A practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, Inc., New York

Bernhardt R (2006) Cytochromes P450 as versatile biocatalysts. J Biotechnol 124:128-45

Berry V, Gascuel O (1996) On the Interpretation of Bootstrap Trees: Appropritate Threshold of Clade Selection and Induced Gain. Mol Biol Evol 13:13

Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 39:519-27

da Fonseca RR, Antunes A, Melo A, Ramos MJ (2007) Structural divergence and adaptive evolution in mammalian cytochromes P450 2C. Gene 387:58-66

Davies L, Williams DR, Aguiar-Santana IA, Pedersen J, Turner PC, Rees HH (2006) Expression and down-regulation of cytochrome P450 genes of the CYP4 family by ecdysteroid agonists in Spodoptera littoralis and Drosophila melanogaster. Insect Biochem Mol Biol 36:801-7

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. . Atlas of Protein Sequence and Structure 5:345–352.

Denison MS, Whitlock JP, Jr. (1995) Xenobiotic-inducible transcription of cytochrome P450 genes. J Biol Chem 270:18175-8

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 43:149-73

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol 20:248-54

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163-7

Felsenstein J (1983) Methods for inferring phylogenies: a statistical view. Numerical taxonomy. Springer-Verlag, Berlin Heidelberg, p 315-334

Felsenstein J (2004) Inferring Phylogenies Sinauer Associates, Inc., Sunderland, Massachusetts

Feyereisen R (2006) Evolution of insect P450. Biochem Soc Trans 34:1252-5

Fujita Y, Ohi H, Murayama N, Saguchi K, Higuchi S (2004) Identification of multiple cytochrome P450 genes belonging to the CYP4 family in Xenopus laevis: cDNA cloning of CYP4F42 and CYP4V4. Comp Biochem Physiol B Biochem Mol Biol 138:129-36

Garfinkel D (1958) Studies on pig liver microsomes. I. Enzymic and pigment composition of different microsomal fractions. Arch Biochem Biophys 77:493-509

Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315-21

Goldstone JV, Goldstone HM, Morrison AM, Tarrant A, Kern SE, Woodin BR, Stegeman JJ (2007) Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family. Mol Biol Evol 24:2619-31

Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jonsson ME, Nelson DR, Stegeman JJ (2010) Identification and developmental expression of the full complement of Cytochrome P450 genes in Zebrafish. BMC Genomics 11:643

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

Graham SE, Peterson JA (1999) How similar are P450s and what can their differences teach us? Arch Biochem Biophys 369:24-9

Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23:1937-45

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500-1

Hardwick JP (2008) Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases. Biochem Pharmacol 75:2263-75

Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure 3:41-62

Hedges SB (1992) The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. Mol Biol Evol 9:366-9

Hoffman SM, Fernandez-Salguero P, Gonzalez FJ, Mohrenweiser HW (1995) Organization and evolution of the cytochrome P450 CYP2A-2B-2F subfamily gene cluster on human chromosome 19. J Mol Evol 41:894-900

Hsu MH, Savas U, Griffin KJ, Johnson EF (2007) Human cytochrome p450 family 4 enzymes: function, genetic variation and regulation. Drug Metab Rev 39:515-38

Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51:673-88

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-5

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82

Lepesheva GI, Waterman MR (2004) CYP51--the omnipotent P450. Mol Cell Endocrinol 215:165-70

Lepesheva GI, Waterman MR (2007) Sterol 14alpha-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. Biochim Biophys Acta 1770:467-77

Li-Hawkins J, Lund EG, Bronson AD, Russell DW (2000) Expression cloning of an oxysterol 7alpha-hydroxylase selective for 24-hydroxycholesterol. J Biol Chem 275:16543-9

Li H, Poulos TL (1996) Conformational dynamics in cytochrome P450-substrate interactions. Biochimie 78:695-9

Li H, Poulos TL (1997) The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. Nat Struct Biol 4:140-6

Li S, Pearl DK, Doss H (2000) Phylogenetic Tree Construction Using Markov Chain Monte Carlo. Jornal of the American statistical Association 95:493-508

Liu Y, Wang J, Liu Y, Zhang H, Xu M, Dai J (2009) Expression of a novel cytochrome P450 4T gene in rare minnow (Gobiocypris rarus) following perfluorooctanoic acid exposure. Comp Biochem Physiol C Toxicol Pharmacol 150:57-64

Mau B, Newton MA, Larget B (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1-12

McArthur AG, Hegelund T, Cox RL, Stegeman JJ, Liljenberg M, Olsson U, Sundberg P, Celander MC (2003) Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. J Mol Evol 57:200-11

Mitschke D, Reichel A, Fricker G, Moenning U (2008) Characterization of cytochrome P450 protein expression along the entire length of the intestine of male and female rats. Drug Metab Dispos 36:1039-45

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8:6-21

Nebert DW, Adesnik M, Coon MJ, Estabrook RW, Gonzalez FJ, Guengerich FP, Gunsalus IC, Johnson EF, Kemper B, Levin W, et al. (1987) The P450 gene superfamily: recommended nomenclature. DNA 6:1-11

Nebert DW, Dalton TP (2006) The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. Nat Rev Cancer 6:947-60

Nelson DR (1998) Metazoan cytochrome P450 evolution. Comp Biochem Physiol C Pharmacol Toxicol Endocrinol 121:15-22

Nelson DR (1999) Cytochrome P450 and the individuality of species. Arch Biochem Biophys 369:1-10

Nelson DR (2003) Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys 409:18-24

Nelson DR (2009) The Cytochrome P450 Homepage. Human Genomics 4:59-65

Nelson DR (2010) Progress in tracing the evolutionary paths of cytochrome P450. Biochim Biophys Acta 1814:14-8

Nelson DR (2010) Progress in tracing the evolutionary paths of cytochrome P450. Biochim Biophys Acta 1814

Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al. (1993) The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. DNA Cell Biol 12:1-51

Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1-18

Nielsen R (2005) Statistical Methods in Molecular Evolution. Springer-Verlag, New York

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581-3

Oleksiak MF, Wu S, Parker C, Karchner SI, Stegeman JJ, Zeldin DC (2000) Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. J Biol Chem 275:2312-21

Oleksiak MF, Wu S, Parker C, Qu W, Cox R, Zeldin DC, Stegeman JJ (2003) Identification and regulation of a new vertebrate cytochrome P450 subfamily, the CYP2Ps, and functional characterization of CYP2P3, a conserved arachidonic acid epoxygenase/19-hydroxylase. Arch Biochem Biophys 411:223-34

Omura T, Sato R (1964) The Carbon Monoxide-Binding Pigment of Liver Microsomes. I. Evidence for Its Hemoprotein Nature. J Biol Chem 239:2370-8

Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1-12

Pevsner J (2009) Bioinformatics and Functional Genomics. Wiley-Blackwell, Hoboken, New Jersey

Pontius JU, Wagner L, Schuler GD (2003) UniGene: A Unified View of the Transcriptome, p 1-12

Poulos TL, Finzel BC, Gunsalus IC, Wagner GC, Kraut J (1985) The 2.6-A crystal structure of Pseudomonas putida cytochrome P-450. J Biol Chem 260:16122-30

Poulos TL, Finzel BC, Howard AJ (1987) High-resolution crystal structure of cytochrome P450cam. J Mol Biol 195:687-700

Reeves JH (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol 35:17-31

Robert FO, Pandhal J, Wright PC (2010) Exploiting cyanobacterial P450 pathways. Curr Opin Microbiol 13:301-6

Ronquist F, Huelsenbeck JP, van der Mark P (2005) MrBayes 3.1 Manual. Draft 5/26/2005

Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, Oxbrow AK, Lewis CJ, Tennant MG, Modi S, Eggleston DS, Chenery RJ, Bridges AM (2006) Crystal structure of human cytochrome P450 2D6. J Biol Chem 281:7614-22

Savas U, Hsu MH, Griffin KJ, Bell DR, Johnson EF (2005) Conditional regulation of the human CYP4X1 and CYP4Z1 genes. Arch Biochem Biophys 436:377-85

Simpson AE (1997) The cytochrome P450 4 (CYP4) family. Gen Pharmacol 28:351-9

Stamatakis A, Ludwig T, Meier H (2005) RAxML-II: a program for sequential, parallel and distributed inference of large phylogenetic trees. Concurrency Comput.: Pract. Exp. 17:1705–1723

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456-63

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3:e67

Wang-Buhler JL, Lee SJ, Chung WG, Stevens JF, Tseng HP, Hseu TH, Hu CH, Westerfield M, Yang YH, Miranda CL, Buhler DR (2005) CYP2K6 from zebrafish (Danio rerio): cloning, mapping, developmental/tissue expression, and aflatoxin B1 activation by baculovirus expressed enzyme. Comp Biochem Physiol C Toxicol Pharmacol 140:207-19

Wang J, Liang P (2003) DigiNorthern, digital expression analysis of query genes based on ESTs. Bioinformatics 19:653-4

Wei BQ, Mikkelsen TS, McKinney MK, Lander ES, Cravatt BF (2006) A second fatty acid amide hydrolase with variable distribution among placental mammals. J Biol Chem 281:36569-78

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9

Whittingham LA, Slikas B, Winkler DW, Sheldon FH (2002) Phylogeny of the tree swallow genus, Tachycineta (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol Phylogenet Evol 22:430-41

Wilcox TP, Zwickl DJ, Heath TA, Hillis DM (2002) Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol Phylogenet Evol 25:361-71

Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE (2000) Microsomal cytochrome P450 2C5: comparison to microbial P450s and unique features. J Inorg Biochem 81:183-90

Wilson JY, McArthur AG, Stegeman JJ (2005) Characterization of a cetacean aromatase (CYP19) and the phylogeny and functional conservation of vertebrate aromatase. Gen Comp Endocrinol 140:74-83

Yan J, Cai Z (2010) Molecular evolution and functional divergence of the cytochrome P450 3 (CYP3) Family in Actinopterygii (ray-finned fish). PLoS One 5:e14276

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396-401

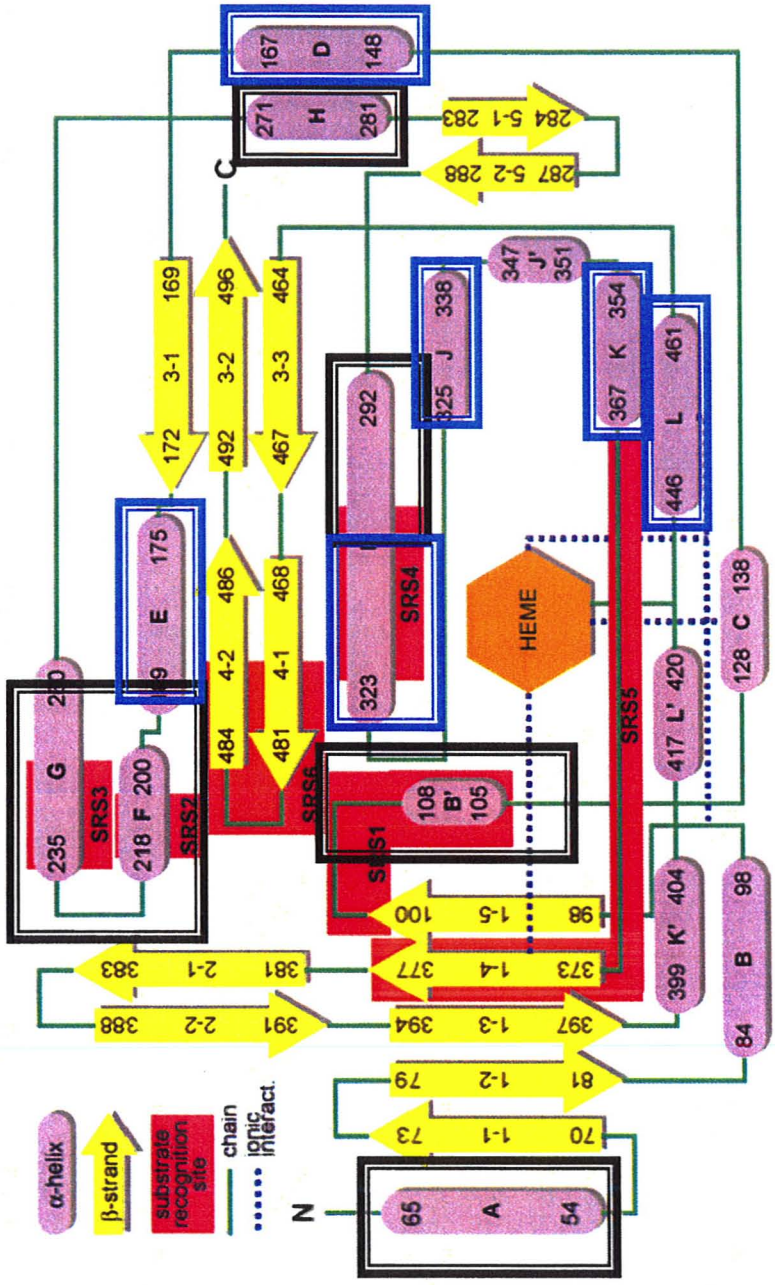
Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14:717-24

Yoshida Y, Aoyama Y, Noshiro M, Gotoh O (2000) Sterol 14-demethylase P450 (CYP51) provides a breakthrough for the discussion on the evolution of cytochrome P450 gene superfamily. Biochem Biophys Res Commun 273:799-804

Zheng Y, Xu D, Gu X (2007) Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits. J Exp Zool B Mol Dev Evol 308:85-96

## 1.5 Figures

**Figure 1.1. A model of the CYP2D6 protein.** A schematic representation of the CYP2D6 protein. This illustration is of the CYP2D6 homology model based on the crystal structure. Important regions of the protein are highlighted including α-helices (purple), β-sheets (yellow), and putative substrate recognition sequences (SRS, red; Gotoh 1992). This figure is from de Graaf *et al.* 2005. Outlined in black boxes are the regions with the highest variability in sequences and which are known to be part of the active site. Those that are outlined in blue are most conserved within the CYP protein structure.


α-helix
β-strand
substrate recognition site
chain
ionic interact.
N
C
HEME
SRS1
SRS2
SRS3
SRS4
SRS5
SRS6

# CHAPTER 2:

# PHYLOGENETIC AND FUNCTIONAL ANALYSIS OF THE VERTEBRATE CYTOCHROME P450 2 FAMILY

Nina Kirischian[1], Andrew G McArthur[2], Caroline Jesuthasan[1], Birgit Krattenmacher [1, 3], and Joanna Y Wilson[1]

[1]Department of Biology, McMaster University, Hamilton, Ontario, Canada.

[2]Andrew McArthur Consulting, Hamilton, Ontario, Canada.

[3] Department of Biology, University of Konstanz, Konstanz, Germany

Hamilton, Ontario, L8S 4K1, Canada

Has been published in Journal of Molecular Evolution 72 (1); 56-71, Copyright © Springer Publisher

**Contributions:** This chapter has been published. Experimental work and data analysis was carried out by N.L.K. under the guidance and supervision of A.G.M and J.Y.W. Annotation of CYP2 sequences in medaka was conducted by B.J. and in stickleback by C.J.

**Abstract**

Cytochrome P450 (CYP) proteins compose a highly diverse superfamily found in all domains of life. These proteins are enzymes involved in metabolism of endogenous and exogenous compounds. In vertebrates, the CYP2 family is one of the largest, most diverse and plays an important role in mammalian drug metabolism. However, there are more than 20 vertebrate CYP2 subfamilies with uncertain evolution and fairly discrete subfamily composition within vertebrate classes, hindering extrapolation of knowledge across subfamilies. To better understand CYP2 diversity, a phylogenetic analysis of 196 CYP2 protein sequences from 16 species was performed using a maximum likelihood approach and Bayesian inference. The analyses included the CYP2 compliment from human, fugu, zebrafish, stickleback, medaka, cow, and dog genomes. Additional sequences were included from rabbit, marsupial, platypus, chicken, frog, and salmonid species. Three CYP2 sequences from tunicate *Ciona intestinalis* were utilized as the outgroup. Results indicate a single ancestral vertebrate CYP2 gene and monophyly of all CYP2 subfamilies. Two subfamilies (CYP2R and CYP2U) pre-date vertebrate diversification, allowing direct comparison across vertebrate classes, while all other subfamilies originated during vertebrate diversification, often within specific vertebrate lineages. Analysis of site-specific evolution indicates that some substrate recognition sites (SRS) previously proposed for CYP genes do not have elevated rates of evolution, suggesting that these regions of the protein are not necessarily important in recognition of CYP2 substrates. Type II functional divergence analysis identified multiple residues in the active site of CYP2F, CYP2A and CYP2B proteins that have undergone radical biochemical changes and may be functionally important.

**Keywords:** Cytochrome P450, vertebrate CYP2 phylogeny, functional divergence, P450 active sites

## 2.1 Introduction

Cytochrome P450 (CYP) enzymes are an ancient superfamily of monooxygenase proteins found in all domains of life (Nelson *et al.* 1993). The CYP superfamily has a total of 977 families of which 69 are present in animals (Nelson 2009). Most vertebrate species have approximately 50-100 CYP genes. Similar CYP families are found across vertebrate species. In rare cases, a CYP family will be found in one vertebrate class but not others; for example, only one family, CYP39, was found in mammals but not in fish (Nelson 2003). The CYP2 family is the largest and most diverse of the vertebrate CYPs (Nelson 2003; Nelson et al. 2004). The number of genes per CYP2 subfamily is variable and can be quite large in some species, particularly mouse, rat and zebrafish. In addition, the CYP2 gene subfamilies are large in number and largely taxonomically distinct (Nelson 2003). Mammalian CYP2s are responsible for the metabolism of structurally diverse drugs, steroids, and carcinogens. In essence, CYP2s play a significant role in the metabolism of a variety of exogenous and endogenous compounds (Lee *et al.* 2008; Lee 2008; Wang and Tompkins 2008). Due to their catalytic importance in Phase 1 oxidation of human pharmaceuticals, CYP2 research has been dominated by studies in mammalian model species and is fundamentally lacking in other vertebrate taxa (Lewis 2002). Mammalian CYP2C and CYP2D genes have attracted the most attention since they metabolize a wide array of metabolic substrates, including many clinically significant pharmaceuticals (e.g. s-warfarin and fluoxetine, respectively).

There are ~12,000 known CYP genes (Nelson 2009) found in all domains of life, and discovery of new CYP sequences is expected due to ongoing genome sequencing of diverse species. Maintaining a clear nomenclature system is essential for gene annotation purposes and to prevent confusion across studies (Nelson 2009; Nelson et al. 2004) . The current nomenclature system is based on a hierarchical clustering of genes into families (>40% amino acid identity) and subfamilies (>55% amino acid identity; Nelson 2003). CYP nomenclature denotes families by number and subfamilies by letter; genes are numbered in order of discovery. For example, the CYP2D6 gene was the 6$^{th}$ gene identified within the 'D' subfamily of the second family. When sequences cannot be clearly placed in a family or subfamily, phylogenetics may assist with nomenclature decisions (Nelson *et al.* 2004). In general, phylogenetic reconstructions support CYP nomenclature in vertebrates.

The CYP2 family includes 29 subfamilies across vertebrate species but only a few subfamilies are found in multiple vertebrate taxonomic classes. Comparison of human and fugu (*Takifugu rubripes*) CYP2 subfamilies found only the CYP2R and CYP2U subfamilies shared amongst a total of 19 CYP2 subfamilies (Nelson 2003). Comparison of CYP2 subfamilies in vertebrate genomes suggests there are lineage-specific CYP2 subfamilies in mammals (CYP 2A, 2B, 2C, 2E, 2F, 2G, 2S, 2W), fish (CYP 2K, 2M, 2N, 2P, 2V, 2X, 2Y, 2Z, 2AA, 2AD, 2AE), birds (CYP2H), and amphibians (CYP2Q; Table 1). The CYP2U and CYP2R genes were present in the vertebrate ancestor and are shared across all vertebrate classes (Nelson 2003). Other CYP2 subfamilies are found in multiple, but not all, vertebrate lineages; CYP2D is

found in the mammalian, amphibian, and avian lineages (Nelson 2009) and CYP2J has been identified in both mammals and amphibians (Nelson 2009). Neither CYP2D nor CYP2J genes are found in actinopterygian (ray-finned fish) species. The CYP2 family has been primarily studied in vertebrate species. However, there are representatives of this family in arthropods and crustacean species, such as CYP2L in lobster (Nelson 1998; Rewitz et al. 2006), and a significant number of CYP2-like genes were found in the sea urchin and *Ciona* genomes 73 and 40+ genes, respectively(Goldstone et al. 2006). Investigating the relationship between vertebrate CYP2 subfamilies is crucial for understanding the evolutionary history of this diverse gene family and may provide clues to the function of novel CYP2 enzymes in non-mammalian lineages.

X-ray crystallography analyses of a *Pseudomonas putida* P450 protein identified 6 putative substrate binding sites (SRS; Poulos *et al.* 1987), regions where amino acids were in close proximity to the substrate and potentially important in substrate recognition and/or binding. Gotoh (1992) aligned mammalian CYP (1, 2, and 3) sequences and determined that regions with high amino acid substitution rates correlated with the *P. putida* binding regions. Gotoh (1992) suggested that these regions were a conserved feature of CYP proteins, that these were substrate recognition sites and that the variable sequence was responsible for the diversity of substrates across the CYP proteins.

Analyses of functional divergence have been completed in a phylogenetic context for both CYP1 (Goldstone *et al.* 2007) and CYP3A proteins (McArthur et al. 2003). These studies have identified a correlation between regions of high rates of divergence and SRSs (Goldstone et al. 2007; McArthur et al. 2003). The analysis of CYP3A genes

revealed some SRSs that correlated with regions of high sequence divergence; however, others were either in regions which presented alignment difficulty or had low sequence divergence (McArthur *et al.* 2003), suggesting that not all of the original 6 SRSs are supported in more robust analyses. Even in cases where the SRS regions were correlated with high sequence divergence, the size and boundaries of the SRSs were not strongly conserved (McArthur *et al.* 2003).

The overall evolutionary relationships of the vertebrate CYP2 subfamilies are poorly understood, which is why the present phylogenetic reconstruction included CYP2 sequences from the major vertebrate lineages: Aves, Amphibia, Actinopterygii, and Mammalia. Species with completed genomes were targeted to ensure coverage of the complete CYP2 complement from multiple species within Mammalia and Actinopterygii. Genome annotation of CYP2 sequences was completed for four species and combined with BLAST searching to amass 196 sequences for phylogenetic reconstruction. CYP2 phylogenetic analyses were performed with maximum likelihood and Bayesian inference to determine the relationships of 24 CYP2 vertebrate subfamilies. Finally, DIVERGE analyses were completed for prediction of type I and type II functional divergence (Gu 2006). Type I functional analysis examined site-specific changes in evolutionary rates for 17 subfamilies and was utilized to predict regions with possible functional divergence within CYP2 proteins. Type II functional analysis detected sites with radical biochemical changes in sister CYP2 subfamilies, with a particular focus on regions of the protein that form the active site.

## 2.2 Methods

### *2.2.1 CYP 2 gene sequences*

CYP2 sequences were acquired via three methods: the Cytochrome P450 homepage (Nelson 2009) BLAST server (http://drnelson.utmem.edu/CytochromeP450.html), *de novo* sequence prediction, and from sequence databases (NCBI Genbank and Ensembl; Table 2). The complete CYP2 gene complement had been previously identified in fugu (*Takifugu rubripes*; Nelson 2003; Nelson 2009), zebrafish (*Danio rerio*; Nelson 2009), and human (Homo sapiens; Nelson 2009; Nelson et al. 2004) and these sequences were retrieved from the Cytochrome P450 homepage BLAST server. These sequences represent fully curated and annotated sequences with nomenclature approved by the P450 nomenclature committee. *De novo* CYP2 sequences were predicted and annotated for the three-spined stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), and dog (*Canis familiaris*) genomes (Supplementary Table 2). CYP2 sequences from fugu were the primary basis of annotation of stickleback (genome assembly version 1; V1) and medaka (V1). Additional sequences from zebrafish CYP2 subfamilies (CYP 2V, 2AA, 2AD, and 2AE) were used in genome searching and annotation because these subfamilies were absent in fugu (Nelson 2003; Nelson 2009). CYP2 sequences for the dog (V2.1) genome were predicted *de novo* using human CYP2 sequences as a reference. CYP2 sequences from the cow (Bos taurus) genome were available on the Cytochrome P450 homepage BLAST server and were used in searching the V4 genome assembly. Lastly, human CYP2 sequences were utilized in exhaustive and extensive BLAST searches (primarily GenBank;

limited searching in Ensembl) to identify additional CYP2 sequences in a number of vertebrate species: rabbit (*Oryctolagus cuniculus),* koala (*Phascolarctos cinereus),* opossum *(Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis* and *X. laevis*), rainbow trout (*Oncorhynchus mykiss*), and Atlantic salmon (*Salmo salar*).

The annotation approach for CYP2 sequences was the same for all *de novo* sequences. Existing sequences (fugu, zebrafish, and human) were utilized in BLAST searches and genome regions with high identity (>60%) were retrieved. Based on previous studies of mammalian CYP2 sequences (Nelson 2003; Nelson et al. 2004) and our own analyses of fugu and zebrafish CYP2 genes (data not shown), CYP2s are composed of nine exons, with the exception of CYP2R and CYP2U genes which have five exons per gene (Rogers and Wall 1980). Exon size is similar for all CYP2 exons in mammalian (Nelson *et al.* 2004) and actinopterygian species (data not shown), but intron size can vary significantly (Nelson 2003). All vertebrate CYP coding transcripts are approximately 1500 base pairs in length. BLAST searches with high sequence similarity and nine (or five if CYP2U and CYP2R) exons found in appropriate order (i.e. exon 1 must be adjacent to exon 2, etc.) were retrieved. These nucleotide sequences were imported into MacClade 4.0 (Maddison and Maddison 2000) for annotation of exon and intron boundaries, based on derived eukaryotic consensus splice sites (Nelson et al. 2004; Rogers and Wall 1980). Putative CYP2 sequences were assessed for accurate splice site boundaries via BLAST 2 Sequence server (i.e. dog CYP2F vs. human CYP2F). For quality assurance, all genes were assessed for the

appropriate number of exons, correct exon order, similar exon sizes, a coding transcript length of ~1500 bp and the presence of start and stop codons. For all *de novo* sequences, RefSeq and EST data were used, where possible, to support the annotated sequences. Putative subfamily assignments were derived based on a sequence identity matrix generated via a CLUSTALX (Thompson *et al.* 1997) algorithm in BioEdit (Hall 1999). All *de novo* annotations were submitted to the P450 nomenclature committee for sequence verification and naming.

*2.2.2 CYP2 gene alignment*

All retrieved and annotated CYP2 amino acid sequences were aligned using CLUSTALX (Thompson *et al.* 1997). *Ciona intestinalis* sequences (CYP2J30, CYP2N, CYP2U1) were used as an outgroup based on the phylogenies of Goldstone *et al.* (2007). Manual adjustments to the alignment (Supplementary CYP2_alignment file) were performed using MacClade 4.08 (Maddison and Maddison 2000) and Mesquite 2.0 (Maddison and Maddison 2009). Sequences that presented difficulties within the alignment were removed. For example, sequences with poor alignment (i.e. <40% amino acid identity to other CYP2s, large gaps in data within the trimmed alignment, or poor alignment in regions of strong homology), partial sequences (i.e. those with less than 450 amino acids or had incorrect number of exons), and identical sequences with different names were removed. Alignment regions with gaps and uncertain homology were excluded (masked) from the final phylogenetic analyses. The

CYP2 family phylogenetic reconstructions were based on 196 aligned sequences with a total length of 326 amino acids (Supplementary CYP2_alignment file).

### *2.2.3 Phylogenetic analyses*

Phylogenetic reconstruction of CYP2 evolutionary history was performed by maximum likelihood and Bayesian inference, both using the JTT+I+Γ (Jones *et al.* 1992) substitution model with unequal amino acid frequencies determined by ProtTest (Abascal *et al.* 2005). The maximum likelihood analyses were computed on a RAxML (Randomized Accelerated Maximum Likelihood) BlackBox server (Stamatakis *et al.* 2008) with 100 bootstrapping replicates. Bootstrapping analysis was computed using maximum likelihood with estimated proportion of invariant sites and with indicated outgroup (*Ciona*) sequences.

Phylogenetic reconstructions with Bayesian inference were completed using the MrBayes computer software program (V3.1.1; Huelsenbeck *et al.* 2001). $MC^3$ (Metropolis-coupled, Markov chain, Monte Carlo) searches were performed using four incrementally heated chains with distinct random initial trees for 20 million generations, with a sampling frequency every 1000 generations. Posterior probabilities were estimated after the removal of $MC^3$ burn-in. Posterior probabilities for nodes that appeared in the maximum-likelihood tree but were not present in the consensus Bayesian phylogenetic reconstruction were assessed using 19.96 million-sampled topologies from the Bayesian analysis in PAUP 4.0 (Swofford, 2000).

Additional maximum likelihood analyses and Bayesian inferences were completed for all CYP2 sequences in two clusters: the CYP2C-2E cluster and the CYP2J-2P-2N-2AD-2AE-2Z cluster. These two clusters had high branch order complexity and poor support within the full phylogenetic analyses (Supplementary Figure 1). The analyses for the CYP2C-2E and CYP2J-2P-2N-2AD-2AE-2Z clusters included a more relaxed masking of the alignment, resulting in alignment length increased by 20 and 17 amino acids, respectively. Phylogenetic reconstructions of the modified alignments were conducted via maximum likelihood analyses.

### *2.2.4 Functional analysis*

To develop a better understanding of CYP2 functional evolution, an analysis of the amino acid alignment in context of the maximum likelihood tree was conducted using DIVERGE (Gu 1999). Type I functional analyses with DIVERGE utilize a phylogenetic tree to assess site-specific changes in evolutionary rates within amino acid alignments when comparing subclades (CYP2 subfamilies in our case). DIVERGE uses the coefficient of evolutionary functional divergence ($\theta$) to measure change in site-specific evolutionary rate: $\theta = 0$ indicates no functional divergence while increasing values indicate increasing functional divergence, with $\theta = 1$ being the maximum. Utilizing the coefficient of evolutionary functional divergence ($\theta$), we tested for significant functional divergence for each of the pairwise comparison of 17 different CYP2 subfamilies (likelihood ratio test (LRT), $p<0.05$).

Type II functional analyses with DIVERGE assessed amino acid positions with physico-chemical changes between subfamilies. We focused on radical changes and applied a cut-off value of $\theta > 1$ for site-specific posterior probabilities. Type II functional analyses were completed for sister subfamilies identified in our phylogenetic analyses (CYP 2C/2E, 2AA/2X, 2K/2W, 2F/2A/2B, 2J/2P/2AD/2N/2Z, and 2U/2R). The sequence alignment was reassessed for each subfamily cluster. Regions of uncertain homology were reexamined and a larger number of residues were included in the type II functional divergence analyses because of the higher similarity of sequences within sister subfamilies as compared to the global CYP2 alignment.

## 2.3 Results

### *2.3.1 CYP2 annotation and nomenclature*

The CYP2 alignment was composed of 196 sequences of which 58 were *de novo* gene annotations, representing 24 CYP2 subfamilies (Figure 1; Table 1). The CYP2 complement for a given vertebrate species ranged from 12 to 20 genes for most species. Some species had large numbers of CYP2 genes (e.g. zebrafish had 44 CYP2 genes; Table 2). There were 27 CYP2 genes for opossum even though this may not be the complete gene complement for this species, since the opossum genome project assembly and annotation were incomplete at the time of this study and we did not exhaustively search the genome. CYP2 gene expansion was not uniform across CYP2 subfamilies but was concentrated in a distinct subset of subfamilies; opossum, for example, had 8 genes

in the CYP2C subfamily, while zebrafish had 12 genes in the CYP2AA subfamily, 9 in the CYP2K subfamily and 7 in the CYP2X subfamily.

In nearly all cases, genome annotation efforts identified sequences from most known mammalian or fish CYP2 subfamilies. A CYP2D gene has not been identified, with confidence, in the dog genome via *de novo* annotation; although there are regions with some high sequence similarity to human CYP2D6, a continuous region containing nine exons was not found in the dog genome (V2.1). Medaka and stickleback genes were not identified for the CYP2V, CYP2AA, and CYP2AD subfamilies, yet these subfamilies are found in zebrafish (Nelson 2009). The genes retrieved for opossum and rabbit most likely account for the majority of CYP2 genes in these species, as representatives from most of the mammalian CYP2 subfamilies (with the exception of CYP2D and CYP2S) were found.

CYP2 sequences were available from the mouse and rat genomes. However, these species have very large (>35 genes per species) CYP2 gene complements due to high gene duplication rates in some subfamilies (Nelson *et al.* 2004). As such, rodent CYP2 genes (including rodent-specific subfamilies CYP2T and CYP2AB) were excluded from our analyses, particularly because there were other mammalian genomes available.

Overall, 5 vertebrate CYP2 subfamilies were not included as they were either not identified by BLAST search (CYP2Q, CYP2AC, CYP2AF) or were rodent-specific (CYP2T, CYP2AB; Nelson 2009; Nelson et al. 2004) . The CYP2Q and CYP2AF

subfamilies are only found in amphibian and avian species, respectively; the CYP2AC subfamily is found in both avian and amphibian species.

### *2.3.2 CYP2 phylogenetic analyses*

For the analysis of overall CYP2 phylogeny, $MC^3$ burn-in length was 40,000 of 20,000,000 generations, resulting in 19,960,000 sampled trees for calculation of posterior probabilities. Replicate $MC^3$ analysis supported convergence via Are We There Yet (AWTY, data not shown; Nylander *et al.* 2008). The overall topology of the maximum likelihood tree was in agreement with the Bayesian inference consensus tree (Supplementary Figure 1). The phylogenetic strength of this analysis was interpreted based on bootstrap values and posterior probabilities. Internal node resolution for CYP2 phylogeny had strong bootstrapping and posterior probability values >88/>0.93, respectively (Figure 2; Figure 3; Supplementary Figure 1). All subfamilies clustered into monophyletic groups, with patterns of internal branching following vertebrate speciation patterns (Li *et al.* 2007; Prasad *et al.* 2008), with few exceptions (CYP2N/CYP2AD, frog CYP2C8, chicken CYP2C45, stickleback CYP2N17, and fugu CYP2N12). Some subfamilies (CYP2C and the cluster including CYP2J) illustrated a complex history of speciation and gene duplication. Gene clusters surrounding mammalian CYP2C (Figure 2) and CYP2J (Figure 3) were subjected to separate phylogenetic analyses.

The placement of some sequences in the CYP2 vertebrate phylogeny did not match their assigned nomenclature. The frog CYP2C8 and chicken CYP2C45 sequences cluster outside of the remaining CYP2C subfamily (Figure 1). Chicken CYP2C45 clustered with the CYP2H (avian) subfamily, whereas frog CYP2C8 did not cluster with any specific subfamily but was placed between the CYP2Y and CYP2M subfamilies (Figure 1). The two unnamed opossum CYP2 sequences (XP_001374840 and XP_001364901) were found to cluster with the CYP2C and CYP2E subfamilies, respectively (Supplementary Figure 1). Stickleback CYP2N17 and fugu 2N12 did not cluster with other CYP2N sequences but instead clustered with the CYP2AD subfamily with strong support (Supplementary Figure 2).

*2.3.3 Basal vertebrate subfamilies (2U, 2R, 2D)*

The vertebrate CYP2 tree topology was rooted with three *Ciona* CYP2 sequences. Two ancestral subfamilies, CYP2R and CYP2U, were at the base of the vertebrate CYP2 phylogeny (Figure 1). While the placement of the CYP2R and CYP2D subfamilies lacked confidence, the basal position of the CYP2U subfamily was well supported (Supplementary Figure 1). The CYP2U subfamily in our analysis included a total of eight sequences from actinopterygian and mammalian species, whereas the CYP2R subfamily included sequences from actinopterygian, avian, and mammalian species. CYP2D gene loss appeared limited to Actinopterygii and possibly Mammalia (marsupial species only) classes. Full length sequences were not found via BLAST search in all vertebrate classes for the CYP 2U, 2R or 2D subfamilies. Specifically, no full length sequence was

identified for CYP2U and CYP2D in Aves, or CYP2U and CYP2R in Amphibia. However, this likely does not reflect gene loss as partial sequences with high sequence identity were found in their respective genomes (data not shown). For example, a partial avian CYP2U sequence with high (60%) sequence identity to the cow CYP2U1 sequence was found but had only three of five exons and was therefore excluded from our analysis. The evolution of vertebrate CYP2s is shown in Figure 1; a more detailed figure, with complete labels for all branches and node support, is found in Supplementary Figure 1.

### *2.3.4 Actinopterygian CYP2 subfamilies cluster with mammalian CYP2J*

After the basal CYP2U, CYP2R and CYP2D subfamilies, a gene duplication event (Node 7, Figure 1) lead to the evolution of two major clusters of CYP2 subfamilies, one which included the CYP2Z, CYP2AE, CYP2V, CYP2P, CYP2AD, CYP2N and CYP2J subfamilies (Figure 3; Supplementary Figure 1). This cluster (Node 8, Figure 1) presented high complexity with poorly supported branch topology in the vertebrate CYP2 phylogeny. Actinopterygian subfamilies and the mammalian CYP2J subfamily were included in this cluster (Figure 3; Supplementary Figure 1). These subfamilies were subjected to separate phylogenetic analyses to provide higher resolution in the phylogenetic tree (Figure 3; Supplementary Figure 2). In this second analysis, the placement of CYP2AE and CYP2V subfamilies within this cluster remained uncertain (both subfamilies are only found in zebrafish). CYP2Z (Node 8,

Figure 1) genes were identified in all four actinopterygian species (medaka, fugu, stickleback, and zebrafish) and formed the basal branch of this cluster with strong phylogenetic support (Figure 3). The remaining subfamilies (CYP 2J, 2P, 2N, 2AD) shared a common ancestral gene duplication event (Figure 1); however, inconsistencies were found in the branch topology for the focused phylogenetic analysis (Figure 3). The CYP2J subfamily clustered with the CYP2P subfamily (Figure 3) and not CYP2N and CYP2AD (Figure 1), and the bootstrap values and posterior probabilities increased to 38% and 0.85, respectively. In both analyses, CYP2N and CYP2AD subfamilies were sister subfamilies with similar phylogenetic support values.

### *2.3.5 Gene expansions in zebrafish CYP2 subfamilies*

The second major cluster that arose from node 7 (Figure 1) included an expansion of actinopterygian genes (CYP 2X-2AA-2K-2Y subfamilies) and the majority of mammalian CYP2 genes (Node 4, Figure 1). The CYP2AA and CYP2X subfamilies both had highly amplified gene numbers in zebrafish; CYP2AA was a zebrafish-specific subfamily, whereas CYP2X was identified in all actinopterygian species. For the CYP2X subfamily, most actinopterygian species had a single CYP2X gene while zebrafish had seven (Supplementary Figure 1). This zebrafish-specific high gene duplication rate was also found in the CYP2K subfamily (Node 6, Figure 1) and these genes were paralogs to the mammalian CYP2W subfamily. In addition, two putative sequences (opossum CYP2 putative (XP_001369607) and frog CYP2 (NP_001037917)) both clustered with the CYP2W subfamily.

### *2.3.6 Mammalian CYP2 subfamilies*

The majority of the mammalian CYP2 sequences were clustered together with the strongly supported basal actinopterygian subfamily CYP2Y (Node 5, Figure 1). While, the CYP2M subfamily was basal to the mammalian CYP2s, it is restricted to salmonid species and support for its placement was not strong (Supplementary Figure 1). There was a major clade within the CYP2 phylogeny that included only sequences from Mammalia and a single Aves subfamily (Node 4, Figure 1). Within this clade, one major cluster (CYP 2A-2G-2B-2S-2F) represents mammalian-specific CYP2 subfamilies. The tree topology placed the 2F subfamily basal to CYP2A, CYP2G, CYP2B, and CYP2S. Subfamilies CYP2A and CYP2G arose from ancestral gene duplication, as did the CYP2B and CYP2S subfamilies (Figure 1). The avian CYP2H subfamily was basal to the mammalian-specific CYP2C and CYP2E subfamilies (Node 3, Figure 1).

Phylogenetic complexity and amplification of CYP2 paralogs was evident in the CYP 2C-2E-2H cluster (Node 3, Figure 1). The CYP2C subfamily was subjected to separate analysis and support for most nodes increased (Figure 2; Supplementary Figure 3). Interestingly, all human CYP2C genes were located in a single clade, while rabbit and cow CYP2C genes were not (Figure 2).

### *2.3.7 Type I functional divergence*

The SRS regions from Gotoh (1992) were mapped on to the unmasked CYP2 alignment (Supplementary CYP2_alignment.nex). Three SRSs (SRS2, SRS3, and SRS6) were located within the masked regions of our alignment. The remaining SRS regions (SRS1, SRS4, and SRS5) were in areas of our alignment used for phylogenetic analyses (Figure 4). Using the masked alignment and the maximum likelihood tree topology, we determined the evolutionary rates of functional divergence of CYP2s using DIVERGE. A total of 17 of 24 subfamilies were included in the DIVERGE type I analyses as these met the required criteria for site-specific divergence analysis (minimum of 4 sequences per subfamily). The coefficient of evolutionary functional divergence *($\theta$)*, its standard error, and the maximum likelihood ratio (LRT) were determined for each pairwise comparison (Supplementary Table 1). A heat map (Figure 4b) was generated based on type I pairwise comparisons of all 17 CYP2 subfamilies in DIVERGE. With the exception of five pairwise comparisons (CYP 2N/2P, 2N/2Z, 2Z/2D, 2N/2Y, 2N/2A), the remaining 131 pairwise comparisons of vertebrate CYP2 subfamilies exhibited statistically significant divergence in site-specific rate of evolution *(LRT, $p<0.05$)*. However, divergence of site-specific rates of evolution was not clustered in the SRS regions or any other part of the alignment, but was distributed across the majority of the CYP2 alignment (Figure 4B).

### *2.3.8 Type II functional divergence*

Additional residues were added to SRS regions in several cases for type II functional analyses because we reassessed the alignment for each sister subfamily

comparison. Four sister subfamily comparisons (2U/2R, 2AD/2N, 2A/2B/2F and 2AA/2X) added residues in SRS2, SRS3, and SRS6. The 2K/2W and 2C/2E subfamily analyses added residues in SRS2 and SRS6. Type II functional analyses of the 2J/2AD/2N/2P/2Z subfamilies did not add any additional residues in the SRS regions. Type II functional divergence analyses identified residues with radical biochemical changes between sister subfamilies, with significant site-specific posterior probabilities ranging from $\theta$=1.00 to 7.11. Seven pairwise comparisons (2C/2E, 2K/2W, 2N/2P, 2N/2Z, 2AD/2P, 2AD/2Z, 2AD/2N) had no detectable type II functional divergence.

In the CYP2 vertebrate basal subfamilies, CYP2U and CYP2R, 70/431 sites had radical changes ($\theta$ >1) but these were not clustered within the protein. There were 8 amino acids with radical changes in SRS3 (helix G), including a site with a high site-specific posterior probability change ($\theta$>5.00). At this site, an R in CYP2U (R288 in human CYP2U1) changes to an F in CYP2R proteins (F247 in human CYP2R1), causing a change from a hydrophobic to a positive amino acid (data not shown). The SRS5 region had 3 radical changes of 10 possible sites.

In the CYP2J/2N/2AD/2P/2Z subfamily cluster (Node 8, Figure 1), radical amino acid changes were only seen between CYP2J and either CYP 2P or 2Z (Table 3). A total of 14 radical changes were dispersed throughout the protein. Interestingly, one site (T196 in human CYP2J2) had radical changes for most CYP2J comparisons, with the exception of CYP 2J/2AD. The threonine (hydrophilic) in the CYP2J proteins changed to valine (hydrophobic) in CYP2N, 2P, and 2Z proteins (Table 3). Only one site had a radical change between CYP2J and CYP2AD (Table 3).

Type II functional divergence analysis detected 100 radical amino acid changes ($\theta$>1) for the CYP 2AA/2X comparison. Most of these sites were distributed throughout the protein, with some clusters in helices A, F, K", the FG-loop, and B4-1 sheet. The FG loop had the highest number of radical changes and contained the sites with the highest $\theta$ (data not shown).

For the CYP 2F/2A, CYP 2F/2B, and CYP 2A/2B comparisons a total of 95, 54, and 36 radical changes ($\theta$>1) were identified, respectively. The largest number of radical changes and highest posterior probability values were found in comparisons with CYP2F (2F/2A > 2F/2B >2A/2B). A total of six radical changes were shared in all 3 comparisons (residue 189, 276, 277, 324, 406 and 476 in Cow_CYP2A13). Most radical changes were detected in the active site (Figure 7). Helix F, which contains SRS2, had changes with the highest $\theta$.

## 2.4 Discussion

### *2.4.1 CYP2 gene sequences*

Two-thirds of the sequences in the CYP2 alignment were from either the P450 homepage or *de novo* annotation (Table 2). These sequences have been verified and named by the P450 nomenclature committee. Species with complete, annotated genomes typically had 12-20 CYP2 genes per genome, although species that are known for high gene amplification can encompass well over 40 genes (e.g. zebrafish). Gene number amplification was detected in the Actinopterygii class for the CYP2X, 2AA, and 2K

subfamilies. These three subfamilies are predominately populated with zebrafish sequences (Supplementary Figure 1). Amplification of CYP2 sequences is thus not uniform but specific to certain subfamilies. Similarly, high gene duplication rates are seen in rodents and marsupials for specific subfamilies (CYP 2B, 2C, 2D, 2J; Nelson *et al.* 2004). These subfamilies are composed of 12-23 genes in the mouse genome and 6-16 genes in the rat genome (Nelson 2009). The CYP2C33v4 cluster is only found in opossum species, although possible CYP2C33v4 orthologs have been identified in pig (Nelson 2009). Functional and expression data is limited for these genes. It is unclear why an expansion of some CYP2 subfamilies occurred in the rat and mouse genomes; perhaps further research can shed light on their expression, function, and specificity (Nelson *et al.* 2004).

This study focused on vertebrate CYP2 genes in species with completed genomes. Representative species from the vertebrate lineages (placental and marsupial mammals, bird, amphibian and fish) were chosen to ensure wide coverage of vertebrates. Due to high amplification of CYP2 genes in rodents (Nelson 2009), we chose to exclude rodents and focus on mammalian species where the number of CYP2 genes were more limited but included the major mammalian CYP2 subfamilies. Zebrafish have a large number of duplications within the CYP2 family yet there are a limited number of genomes completed within the Actinopterygii lineage to compare and thus zebrafish sequences were included. The full CYP2 complement was thus obtained for 3 mammalian (human, dog, cow) and 4 actinopterygian (zebrafish, fugu, medaka, stickleback) species. A significant number of CYP2 genes, representing most mammalian CYP2 subfamilies,

were included from rabbit and opossum. Chicken and frog species were included to provide coverage of the major branches of vertebrate diversity.

Five vertebrate CYP2 subfamilies (CYP2T, CYP2Q, CYP2AB, CYP2AC, and CYP2AF) were not identified during BLAST searching, which utilized human and fugu CYP2 sequences. The CYP2T, CYP2Q, CYP2AB, and CYP2AC subfamilies are specific to rodents, amphibian, avian and rodent/amphibian species, respectively (Nelson 2003; Nelson et al. 2004). A CYP2G sequence from dog has also been recently identified (Nelson 2009) but was not found by our search strategy. Thus, our sequences are not a complete representation of every possible CYP2 gene or subfamily in vertebrates but an assessment of CYP2 genes in the major subfamilies, particularly in the mammalian and actinopterygian lineages.

### *2.4.2 CYP2 nomenclature issues*

The phylogenetic reconstructions did not support all CYP2 gene nomenclature and we suggest that the nomenclature be reassessed for some genes. The frog CYP2C8 (NP_001079610) sequence clustered outside of the CYP2C subfamily and thus the appropriate subfamily for this gene is unclear. The chicken CYP2C45 (NP_001001752) clustered well within the CYP2H subfamily. Certain sequences, including opossum CYP2 (XP 001374840), lacked nomenclature at retrieval but our analyses found the gene to cluster with the CYP2C subfamily. Similarly, our analyses strongly support the placement of the opossum CYP2 (XP 001364901) gene within the CYP2E subfamily.

Surprisingly, both stickleback CYP2N17 and fugu CYP2N12 may need reassignment to the CYP2AD subfamily.

### *2.4.3 Cytochrome P450 2 family phylogeny*

Our CYP2 phylogenetic analysis generated a high resolution tree with strong support (posterior probabilities and bootstrap values) for all internal nodes. With the exception of possible nomenclature problems for the CYP2N and CYP2AD subfamilies, all CYP2 subfamilies were monophyletic. The branching patterns within subfamilies often matched vertebrate speciation patterns (Prasad *et al.* 2007). For the Actinopterygii, speciation patterns had zebrafish as the most distal, medaka as an intermediate and fugu and stickleback as sister species, as expected (Li et al. 2007).

### *2.4.4 Basal Vertebrate Subfamilies (2U, 2R, 2D)*

Contrary to that suggested by Nelson *et al.* (2004), the CYP2 phylogeny indicates that the vertebrate ancestor had 3 CYP2 genes; a CYP2U, CYP2R, and another CYP2 gene that diversified into the remaining CYP2 subfamilies (Figure 5). Prior CYP phylogenetic analyses showed CYP2U and CYP2R subfamilies to have deep branches within the CYP2 family (Nelson *et al.* 2004). While support for the basal position of these subfamilies is strong, CYP2R and CYP2D placement within the basal branches is weak in our phylogenetic tree. However, CYP2R and CYP2U are the only CYP2

subfamilies with representatives from all vertebrate classes (Figure 5) and the primary sequence structure of both CYP2U and CYP2R includes five exons; all other CYP2 genes, including CYP2D genes, contain nine exons (Nelson *et al.* 2004). Collectively, this supports the placement of CYP2U as the ancestral subfamily at the base of the CYP2 phylogenetic tree followed by CYP2R. With the development of the CYP2D ancestor, the gene structure for CYP2 genes increased to nine exons (2004). In our analyses, representatives from CYP2U, CYP2R and CYP2D were not identified from all vertebrate lineages yet CYP2U and CYP2R are expected in all vertebrate taxa, as seen in Thomas (2007) . CYP2D sequences are expected in all but Actinopterygii lineages (Nelson 2003; Thomas 2007). Determining gene loss in certain lineages can be difficult based on available sequences and genomes. For example, CYP2S, CYP2F, CYP2R and CYP2U genes have yet to be identified in rabbit (Nelson 2009), for which genome sequences are lacking. Partial sequences for avian CYP2U and CYP2D, partial amphibian CYP2U, and a full amphibian CYP2R sequence have been identified by Thomas (2007), which were considered for our analysis of gene duplication history (Figure 5). We hypothesize that the lack of CYP2D sequences in the Actinopterygii class is the product of gene loss (Figure 5) based on their lack in all four actinopterygian species with completed genomes. Gene duplication and speciation patterns were also inferred for the CYP2AA/2X/2K/2W and CYP2Y/2M/2F/2S/2B/2G/2A clusters and are in supplementary Figures 4 and 5, respectively.

### *2.4.5 Actinopterygian CYP2 subfamilies cluster with mammalian CYP2J*

For much of the phylogenetic tree, Actinopterygii and Mammalia CYP2 subfamilies were located in distinct clades (Figure 1). One major cluster of actinopterygian CYP2 subfamilies (2P-2N-2AD-2AE-2V-2Z) contained the mammalian and avian CYP2J subfamily (Node 8, Figure 1). This clade was subjected to separate phylogenetic analysis to accurately resolve complexity associated with internal nodes (Figure 3; inferred speciation and duplication patters are in supplementary Figure 6). The CYP2J subfamily shares a common ancestral CYP2 with the CYP2P subfamily (Figure 3), even though CYP2P genes have higher sequence similarity to CYP2N and CYP2AD (data not shown). Studies by Oleksiak *et al* (2003) suggest that CYP2J, CYP2N and CYP2P enzymes have functional similarities; the regio- and enantioselectivities of killifish (*Fundulus heteroclitus*) CYP2P3 for arachidonic acid were similar to the mammalian CYP2J2 gene. The CYP2J, CYP2P and CYP2N enzymes have similar structure, metabolic pathways, and expression patterns, with high levels in heart, kidney, and intestines (Scarborough *et al.* 1999). Oleksiak *et al.*(2003) suggested that CYP2P and CYP2N had a common ancestral gene, which is supported in our phylogenetic study with inclusion of CYP2J and CYP2AD in that common ancestry. Further investigation into subfamily CYP2AD would be of importance in identifying what functional role it may have in common with CYP2J, CYP2P, or CYP2N. The CYP2AE and CYP2V subfamilies were only represented by a single zebrafish gene and their exact placement relative to the CYP2AD, CYP2N, CYP2P, CYP2J, and CYP2Z subfamilies was unresolved (Figure 3).

### *2.4.6 Mammalian CYP2C and CYP2E subfamilies*

The most recently evolved subfamilies in CYP2 vertebrate evolution are the mammalian CYP2C and CYP2E subfamilies (Figure 6). A series of gene duplication events occurred within the CYP2C subfamily, with the first giving rise to the CYP2C and CYP2C33v4 clades (Node 1, Figure 1). The CYP2C33v4 clade is basal and includes many marsupial genes from both opossum and koala. No CYP2C genes from placental mammals clustered with the CYP2C33v4 genes, suggesting gene loss in most, but not all, placental mammals (Figure 6) as a CYP2C33v4 gene has been identified in pigs (Nelson 2009). Within the CYP2C33v4 cluster, two koala CYP2C sequences (CYP2C47 and CYP2C48) have high sequence similarity (>80%) to four CYP2C33v4 sequences (data not shown).

The CYP2C clade contained both placental and marsupial genes, although these were distinctly clustered and separated by a speciation event (Figure 6). The CYP2C sequences are known to have LINE-1 (L1) elements, which are interspersed repetitive DNA elements that can replicate via retrotransposition (Boissinot et al. 2000; Nelson et al. 2004). The L1 elements are primarily found in intron 5 of CYP2C sequences. Multiple L1 elements are found in CYP2C18, CYP2C19, and CYP2C9 (Nelson *et al.* 2004), which have sequence similarity of 84%-92%. These types of recombination factors can lead to production of novel hybrid CYP2C genes.

### *2.4.7 Type I functional divergence*

Six substrate recognitions sites (SRSs) were identified by a correlation between high substitution rates and similarity in location to residues that appeared to be involved in substrate interactions in a *P. putida* P450 101A gene (Gotoh 1992). The SRSs were identified based on mammalian CYP sequence, including CYP2A, CYP2B, CYP2C, CYP2D, CYP2E and CYP2F genes, and were suggested to be functionally significant for all CYPs (Gotoh 1992; Lewis 2003). Gotoh (1992) suggested residues with high rates of change are responsible for the specificity and functional diversity of CYPs. Pairwise comparisons of 17 CYP2 subfamilies were completed using DIVERGE to determine which regions of the CYP2 genes had elevated rates of evolutionary divergence. The six SRS regions proposed by Gotoh (1992) were mapped onto the CYP2 alignment (Figure 4a); three of these SRSs were in masked regions of the alignment and three SRSs were included in our phylogenetic and divergence analyses. There was no association between high evolutionary rates and the SRS locations in the alignment (Figure 4b). The heatmap (Figure 4b) illustrates that there are residues with high amino acid divergence; however, with the exception of selected regions of $\theta<0.5$, the divergence was distributed throughout the protein alignment. Notably, SRS1 showed no statistically significant change in evolutionary rates of functional divergence (Figure 4).

Analysis of subfamily divergence indicates some subfamily-specific patterns (Figure 4b). For example, comparisons involving subfamilies CYP2A or CYP2E have high evolutionary rates of functional divergence throughout the majority of the sequence alignment, whereas comparisons to CYP2N show low rates of divergence. A total of five

subfamily comparisons were not statistically significant based on LRT values (Supplementary Table 1). One such comparison was for CYP2N/CYP2P. These two subfamilies seem to share common functional characteristics (Oleksiak *et al.* 2003), supported by our DIVERGE results. The CYP2N/CYP2A comparison also lacked statistically significant divergence, which was surprising due to the large evolutionary distance between the two subfamilies (Figure 1). Further examination of these two subfamilies would be of interest to determine whether CYP2N (fish) and CYP2A (mammalian) share commonalities in function.

### *2.4.8 Type II functional divergence*

Our type II functional analyses identified sites with radical biochemical changes between closely related subfamilies. These sites were localized to the appropriate protein structure (helix or B-sheet) and clusters of sites were identified in the amino acid sequence. Particular attention was paid to those sites found in regions important to the active site of the protein. Like the type I functional analyses, type II functional analyses of CYP2 subfamilies had biochemical changes spread throughout the protein without strong clustering of radical changes for most subfamilies. Seven of the pairwise subfamily comparisons had no type II functional divergence, indicating that the CYP 2C/2E, 2K/2W, 2N/2P, 2N/2Z, 2AD/2P, 2AD/2Z, 2AD/2N sister clades may be functionally similar. Certainly, functional studies (Oleksiak et al. 2003) and type I divergence support similar function between CYP2J and CYP2P subfamilies.

Studies of cytochrome P450 crystal structure have identified functional regions of the protein. Conservative sites appear to surround the heme group (de Graaf et al. 2005; Hasemann et al. 1995; Mestres 2005) and conformational similarity was found in eukaryotic and bacterial P450s for this region, even though overall sequence identity was low at 10%-30% (de Graaf et al. 2005; Hasemann et al. 1995; Mestres 2005). The conserved interior structure extends to helices D, E, J, J', K, K', K'', L, and the central-carboxyl terminal of the I helix (de Graaf et al. 2005; Hasemann et al. 1995; Mestres 2005). Higher variability in structure was associated with sites involved in substrate recruitment and binding (Hasemann et al. 1995; Mestres 2005). The active site has been associated with helices B' (SRS1), C, C', F (SRS2), G (SRS3), the N-terminus of the I helix, and the B1-4 and B4-2 (SRS6) sheets (de Graaf et al. 2005; Hasemann et al. 1995; Lewis 2002; Rowland et al. 2006). Helix F and G and the FG-loop have the highest variability in sequence identity and length due their importance for regiospecificity and spatial movement (Mestres 2005).

Type II functional analysis identified a limited number of radical changes in the CYP 2J/2P and CYP2J/2N comparisons (Table 3), supporting Oleksiak et al.'s (2003) contention that CYP2J, 2N and 2P have similar catalytic function. CYP2J and CYP2P metabolize arachidonic acid and share regio- and enantioselectivities (Oleksiak et al. 2003). CYP2P and CYP2J have not only high sequence similarity, but similar catalytic function and tissue distribution (Oleksiak et al. 2003). Only two sites were identified with radical changes between the CYP2J and CYP2P subfamilies, while a total of 9 radical

changes were detected between the CYP2J and CYP2N subfamilies. The two CYP2J/2P sites were also identified in the type II functional analysis of CYP2J and CYP2N (Table 3). CYP2Ns oxidize arachidonic acid with regio- and enantio-selectivity that are very different from CYP2P (Oleksiak et al. 2000). All radical changes were located within conserved helices or in-between conserved regions, with no sites identified in regions important for the active site.

In contrast to the CYP2J/2N/2P comparisons, type II functional divergence was identified in regions of the active site for CYP2F, CYP2A and CYP2B. Helix F and the B4-2 sheet, which include SRS2 and SRS6 regions, respectively, had the highest site-specific $\theta$ and number of radical changes (Figure 7A). These radical changes were highest between CYP2F and CYP2A subfamilies, suggesting that the substrate access channel for substrate binding may be different between these two subfamilies. During metabolism, the product exits the CYP protein through a region defined by helices G and I; one-third of the residues in this region had radical changes ($\theta > 1$; Figure 7D) between CYP2F and CYP2A. In addition, flexibility and variability in the B' helix, BC loop and SRS6 is necessary for easy passage of the substrate and product (de Graaf et al. 2005; Hasemann et al. 1995). CYP2A and CYP2F type II functional analysis identified one-third of these regions to have radical changes (Figure 7B, 7C). Collectively, this suggests that the substrates for CYP2F and CYP2A proteins are likely to be distinct. For example, CYP2A6 shows high metabolic activity for coumarin 7-hydroxylase (Yano et al. 2006), whereas CYP2F1 are found to be important in metabolizing different pneumotoxins such as naphthalene (Tournel et al. 2007). CYP2A6 and CYP2A13 show similar substrate

selectivity, yet differences are seen in metabolism of substrates (DeVore et al. 2008).

## 2.5 Conclusion

Our phylogenetic reconstruction provides an evolutionary understanding of CYP2 function and diversity. The topology illustrates that most subfamilies are lineage-specific, with the exception of a few ancestral subfamilies (CYP2U, CYP2R, CYP2D; Figure 5). The CYP2 family has diversified throughout all vertebrate species and has expanded gene copy numbers within particular subfamilies (e.g. zebrafish CYP2X, CYP2AA and CYP2K). Since knowledge of function for genes outside of the mammals is limited, our comparative approach provides insight on CYP2 function and diversity for other vertebrate lineages. Our data indicates that the predicted SRSs do not correlate with evolutionary rates of amino acid divergence within the CYP2 alignment. Certain SRSs had small clusters of residues with radical biochemical changes between sister subfamilies but this result was not consistent across all subfamilies nor were the majority of residues in an SRS region involved. Collectively, our type I and II functional divergence suggests that SRSs are not necessarily functionally important for CYP2 proteins. Overall, our evolutionary investigation of CYP2s provides a number of hypotheses worth testing in a functional context.

**Acknowledgements**

We would like to thank Dr. David Nelson (University of Tennessee) for assigning nomenclature to our *de novo* gene annotations and Drs. Brian Golding and Jonathan Stone (McMaster University) for computer cluster access to run our Bayesian analyses and access to PAUP* software for posterior probability node analysis, respectively. We are thankful to Emily Smith and Dr. Golding for helpful comments and suggestions during manuscript revision. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant #328204 to JYW). The Department of Biology, McMaster University provided partial support for N.K.

## 2.6 References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-5

Boissinot S, Chevret P, Furano AV (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 17:915-28

de Graaf C, Vermeulen NP, Feenstra KA (2005) Cytochrome p450 in silico: an integrative modeling approach. J Med Chem 48:2725-55

DeVore NM, Smith BD, Urban MJ, Scott EE (2008) Key residues controlling phenacetin metabolism by human cytochrome P450 2A enzymes. Drug Metab Dispos 36:2582-90

Goldstone JV, Goldstone HM, Morrison AM, Tarrant A, Kern SE, Woodin BR, Stegeman JJ (2007) Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family. Mol Biol Evol 24:2619-31

Goldstone JV, Hamdoun A, Cole BJ, Howard-Ashby M, Nebert DW, Scally M, Dean M, Epel D, Hahn ME, Stegeman JJ (2006) The chemical defensome: environmental sensing and response genes in the Strongylocentrotus purpuratus genome. Dev Biol 300:366-84

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

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664-74

Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23:1937-45

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 95-98

Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure 3:41-62

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82

Lee HS, Park EJ, Ji HY, Kim SY, Im GJ, Lee SM, Jang IJ (2008) Identification of cytochrome P450 enzymes responsible for N -dealkylation of a new oral erectogenic, mirodenafil. Xenobiotica 38:21-33

Lee TS (2008) Reverse Conservation Analysis Reveals the Specificity Determining Residues of Cytochrome P450 Family 2 (CYP 2). Evol Bioinform Online 4:7-16

Lewis DF (2002) Homology modelling of human CYP2 family enzymes based on the CYP2C5 crystal structure. Xenobiotica 32:305-23

Lewis DF (2003) Human cytochromes P450 associated with the phase 1 metabolism of drugs and other xenobiotics: a compilation of substrates and inhibitors of the CYP1, CYP2 and CYP3 families. Curr Med Chem 10:1955-72

Li C, Orti G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol 7:44

Maddison DR, Maddison WP (2000) MacClade version 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland Massachusetts

Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. In: 2.72 V (ed)

McArthur AG, Hegelund T, Cox RL, Stegeman JJ, Liljenberg M, Olsson U, Sundberg P, Celander MC (2003) Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. J Mol Evol 57:200-11

Mestres J (2005) Structure conservation in cytochromes P450. Proteins 58:596-609

Nelson DR (1998) Metazoan cytochrome P450 evolution. Comp Biochem Physiol C Pharmacol Toxicol Endocrinol 121:15-22

Nelson DR (2003) Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys 409:18-24

Nelson DR (2009) The Cytochrome P450 Homepage. Human Genomics 4:59-65

Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al. (1993) The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. DNA Cell Biol 12:1-51

Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1-18

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581-3

Oleksiak MF, Wu S, Parker C, Karchner SI, Stegeman JJ, Zeldin DC (2000) Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. J Biol Chem 275:2312-21

Oleksiak MF, Wu S, Parker C, Qu W, Cox R, Zeldin DC, Stegeman JJ (2003) Identification and regulation of a new vertebrate cytochrome P450 subfamily, the CYP2Ps, and functional characterization of CYP2P3, a conserved arachidonic acid epoxygenase/19-hydroxylase. Arch Biochem Biophys 411:223-34

Poulos TL, Finzel BC, Howard AJ (1987) High-resolution crystal structure of cytochrome P450cam. J Mol Biol 195:687-700

Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25:1795-808

Prasad JC, Goldstone JV, Camacho CJ, Vajda S, Stegeman JJ (2007) Ensemble modeling of substrate binding to cytochromes P450: analysis of catalytic differences between CYP1A orthologs. Biochemistry 46:2640-54

Rewitz KF, Styrishave B, Lobner-Olsen A, Andersen O (2006) Marine invertebrate cytochrome P450: emerging insights from vertebrate and insects analogies. Comp Biochem Physiol C Toxicol Pharmacol 143:363-81

Rogers J, Wall R (1980) A mechanism for RNA splicing. Proc Natl Acad Sci U S A 77:1877-9

Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, Oxbrow AK, Lewis CJ, Tennant MG, Modi S, Eggleston DS, Chenery RJ, Bridges AM (2006) Crystal structure of human cytochrome P450 2D6. J Biol Chem 281:7614-22

Scarborough PE, Ma J, Qu W, Zeldin DC (1999) P450 subfamily CYP2J and their role in the bioactivation of arachidonic acid in extrahepatic tissues. Drug Metab Rev 31:205-34

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758-71

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3:e67

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876-82

Tournel G, Cauffiez C, Billaut-Laden I, Allorge D, Chevalier D, Bonnifet F, Mensier E, Lafitte JJ, Lhermitte M, Broly F, Lo-Guidice JM (2007) Molecular analysis of the CYP2F1 gene: identification of a frequent non-functional allelic variant. Mutat Res 617:79-89

Wang H, Tompkins LM (2008) CYP2B6: new insights into a historically overlooked cytochrome P450 isozyme. Curr Drug Metab 9:598-610

Yano JK, Denton TT, Cerny MA, Zhang X, Johnson EF, Cashman JR (2006) Synthetic inhibitors of cytochrome P-450 2A6: inhibitory activity, difference spectra, mechanism of inhibition, and protein cocrystallization. J Med Chem 49:6987-7001

**Table 2.1. Lineage specificity of the vertebrate CYP2 subfamilies.** The 24 CYP2 subfamilies included in this study are shown according to the vertebrate classes in which they are found. Four Actinopterygii (fugu, medaka, stickleback, zebrafish) and three mammalian (dog, cow, and human) species with complete genome sequence availability were included.

| Vertebrate Class | CYP2 Subfamilies | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2A | 2B | 2C | 2D | 2E | 2F | 2G | 2H | 2J | 2K | 2L | 2M | 2N | 2P | 2R | 2S | 2U | 2V | 2W | 2X | 2Y | 2Z | 2AA | 2AD | 2AE |
| **Amphibia**[a] | | | [b] | 2D | | | | | | | | | | | | | | | | | | | | | |
| **Aves** | | | | | | | | 2H | 2J | | | | | | 2R | | | | | | | | | | |
| **Mammalia** | 2A | 2B | 2C | 2D | 2E | 2F | 2G | | 2J | | | | | | 2R | 2S | 2U | | 2W | | | | | | |
| **Actinopterygii** | | | | | | | | | | 2K | 2L | 2M | 2N | 2P | 2R | | 2U | 2V | | 2X | 2Y | 2Z | 2AA | 2AD[c] | 2AE[c] |

[a] Frog CYP2 (NP_001037817) sequence did not cluster with any particular subfamily. It clustered with opossum CYP2 putative (XP_001369607) and they share a common ancestral duplication event with CYP2K and CYP2W. Due to this complexity, no putative subfamily annotation has been proposed for the frog CYP2 (NP_001037817) sequence.

[b] Frog CYP2C8 (NP_001079610) sequence did not cluster with CYP2C subfamily. Frog CYP2C8 clusters between CYP2Y and CYP2M subfamilies (Actinopterygii). Due to this complexity, frog CYP2C8 (NP_001079610) was not included in the 2C subfamily count.

[c] These sequences were only found in the zebrafish genome.

**Table 2.2. CYP2 sequences by source and species.** Sequences were collected during de novo annotation of complete genomes (de novo) or retrieved from the P450 homepage and GenBank. Species with de novo sequences or from the P450 homepage represent the species with complete CYP2 complements. GenBank sequences were retrieved with exhaustive BLAST searching and do not represent all subfamilies for the species.

| Class | Species | *de novo* | P450 Homepage[a] | GenBank |
|---|---|---|---|---|
| *Amphibia* | Frog (*Xenopus tropicalis* | | | 5 |
| *Aves* | Chicken (*Gallus* | | | 8 [b] |
| *Mammalia* | Cow | 20[c] | | |
| | Dog | 12 | | |
| | Human | | 16 | |
| | Koala | | | 2 |
| | Opossum | | | 27[d] |
| | Platypus | | | 1 |
| | Rabbit | | | 17[d] |
| *Actinopterygii* | Atlantic Salmon | | | 1 |
| | Fugu | | 14 | |
| | Medaka | 13 | | |
| | Rainbow trout | | | 1 |
| | Stickleback | 13 | | |
| | Zebrafish | | 44 | |
| *Ascidiacea* | Ciona | | | 3 |

[a] Nelson, DR (2009)

[b] Sequences retrieved from GenBank and Ensembl.

[c] Cow sequences were retrieved from the P450 homepage and used to annotate genome version 4.

[d] Most species have a CYP2 family complement of approximately 15 sequences. The total number of retrieved sequences for opossum and rabbit (27 and 17, respectively) from GenBank suggest that they may be the full CYP2 complement.

**Table 2.3. Sites of radical change in the CYP 2J/2AD/2N/2P/2Z subfamilies.** Sites of radical change were detected in pairwise type II functional analyses of CYP2J, CYP2AD, CYP2N, CYP2P, and CYP2Z subfamilies. The residue positions are provided based on a reference sequence, human CYP2J2. The amino acid changes between subfamilies and the type of change are provided.

| Residue position (Human CYP2J2) | CYP 2J/2N | CYP 2J/2AD | CYP 2J/2P | CYP 2J/2Z | CYP 2P/2Z | Helix or SRS | Amino Acid Changes | Property Changes [1] |
|---|---|---|---|---|---|---|---|---|
| 87 [2](498)[3] | | | | | 2.18 | NA | R to W | +/Hydrophobic |
| 95 [506] | 1.12 | | | | | B | P to K[a] | Hydrophilic/+ |
| 158 [808] | 1.13 | | | | | D | R to S/Y[b] | +/Hydrophilic & Hydrophobic[b] |
| 160 [810] | | | | 1.39 | | D | Q to L/T[c] | Hydrophilic/Hydrophobic & Hydrophilic[c] |
| 166 [816] | | | | 1.34 | | D | L to C/A[d] | Hydrophobic/Hydrophilic & Hydrophobic[d] |
| 181 [844] | 1.09 | | | | | D' | H to V[e]/A[e]/T[f] | +/Hydrophobic[e] & Hydrophilic[f] |
| 196 [860] | 4.01 | | 2.84 | 3.28 | | E | T to V | Hydrophilic & -/Hydrophobic |
| 356 [1084] | 1.80 | | | | | NA | A/E to A | Hydrophobic & -/ Hydrophobic |
| 370 [1110] | 1.20 | | | | | SRS5 | V/A[g] to G/A[h] | Hydrophobic[g]/Hydrophilic & Hydrophobic[h] |
| 384 [1114] | | | | | 1.00 | NA | T/A to T | Hydrophilic & Hydrophobic/Hydrophilic |
| 391 [1122] | 1.78 | | | | | NA | A to G/R | Hydrophobic/Hydrophilic & + |
| | | 1.56 | | | | NA | A to G | Hydrophobic/Hydrophilic |
| | | | | 1.44 | | NA | A to G/E | Hydrophobic/Hydrophilic & - |
| 428 [1159] | | | | | 1.00 | NA | D to N/D | -/Hydrophilic & - |
| 431 [1163] | 1.19 | | | | | NA | Q/E[i] to K/R | Hydrophilic & -/+ |
| | | | 1.24 | | | NA | Q/E[i] to K | Hydrophilic &- /+ |
| 463 [1198] | 4.01 | | | | | L | T to V | Hydrophilic/Hydrophobic |

[1]-Property changes are given for the first/second subfamily in the analyses, + is positively charged, - is negatively charged.

[2]-Residue position based on a reference sequence, Human CYP2J2.

[3]-Indicates the position in the full CYP2 alignment (see supplementary nexus file).

[4]-NA - not applicable, not found in a specific helix or SRS region.

[a]-1Q and 1S amino acid in in chicken 2J2 and opossum 2J2 sequence, respectively.

[b]-Amino acid restricted to medaka CYP2N sequences

[c]-Amino acid found in fugu_CYP2Z2 sequence

[d]-Amino acid found only in stickleback_2Z4 sequence

[e] and [f]-V and A appeared in equal frequency in CYP2N genes, whereas T is only found in zebrafish_CYP2N13 sequence

[g]-A amino acid is only specific to two opossum sequences

[h]-A is only found in fugu_CYP2N11

[i]-E amino acid is found in only one sequence, opossum CYP2J2.

## 2.8 Figures

**Figure 2.1. Vertebrate CYP2 phylogenetic analysis**. A circular phylogenetic representation of the vertebrate CYP2 family with colours highlighting lineage containing in each subfamily. The PROTTEST program determined a JTT+I+G (Jones Taylor Thornton + invariant sites + gamma distribution) model for these protein sequences. A maximum likelihood approach (RAxML) was applied to the final alignment of CYP2 sequences to determine the phylogenetic history. Bayesian inference was computed based on the final CYP2 alignment. Major distal nodes are labeled (1-8) for reference throughout the text and subsequent figures; bootstrapping/posterior probability values were 100/1.00, 98/1.00, 96/1.00, 36/1.00, 99/1.0, 92/1.00, 40/0.95, 91/1.00, for nodes 1-8 respectively. CYP2AE/2V cluster represents both subfamilies that contain one sequence per subfamily.

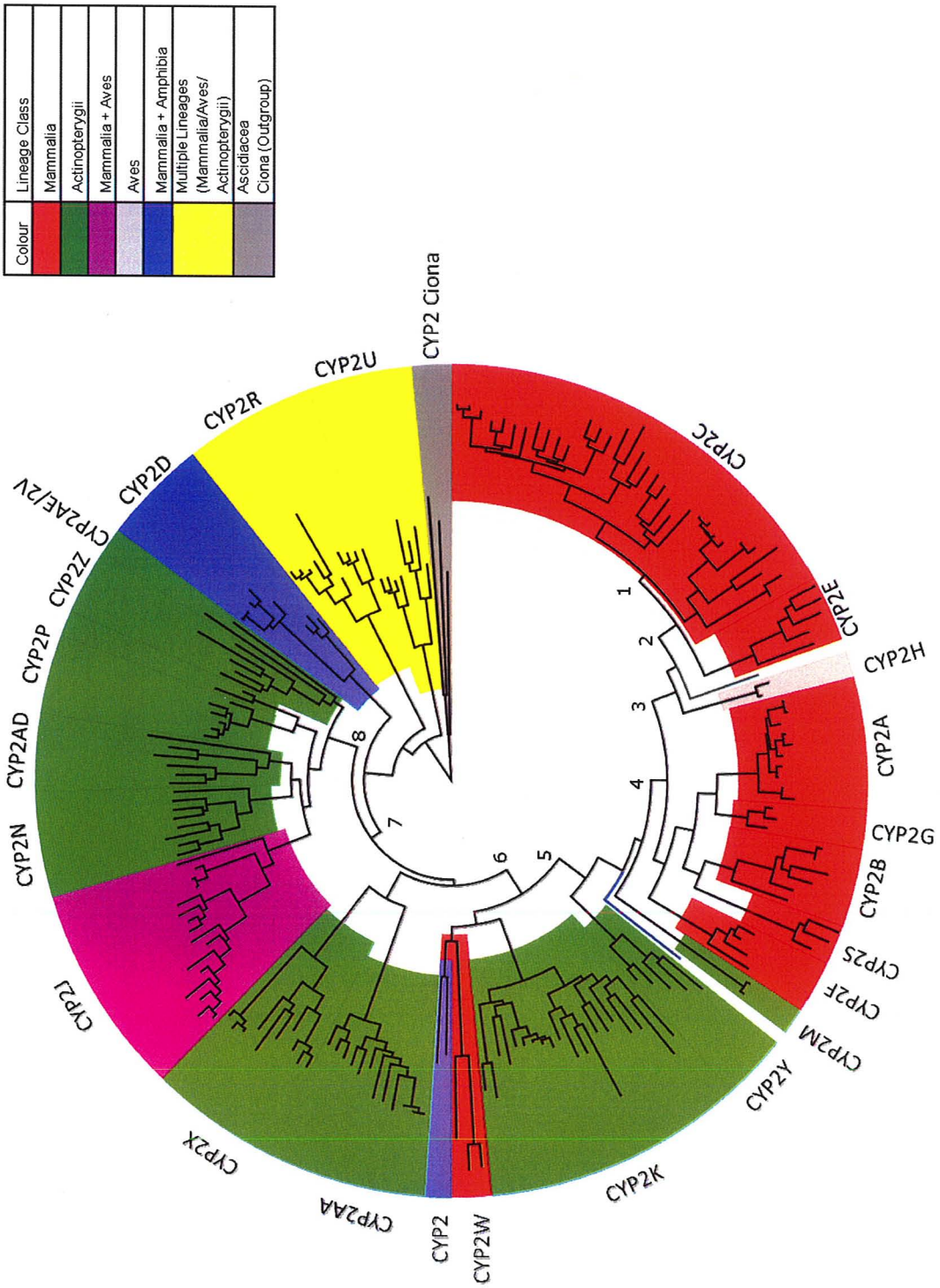

Colour
Lineage Class
Mammalia
Actinopterygii
Mammalia + Aves
Aves
Mammalia + Amphibia
Multiple Lineages (Mammalia/Aves/ Actinopterygii)
Ascidiacea
Ciona (Outgroup)
CYP2 Ciona
CYP2U
CYP2R
CYP2D
CYP2AE/2V
CYP2Z
CYP2P
CYP2AD
CYP2N
CYP2J
CYP2X
CYP2AA
CYP2
CYP2W
CYP2K
CYP2Y
CYP2M
CYP2F
CYP2S
CYP2B
CYP2G
CYP2A
CYP2H
CYP2E
CYP2C
1
2
3
4
5
6
7
8

**Figure 2.2. Phylogenetic analysis of the CYP2C clade.** Phylogenetic analyses of the CYP2C cluster were conducted due to high complexity within the vertebrate CYP2 phylogeny (Node 2, Figure 1). A maximum likelihood tree is shown with bootstrapping and posterior probability values labeled on all distal nodes (See Materials and Methods for details). Placental, marsupial, and outgroup (CYP2E) clades are shaded black, grey, and white, respectively. Internal nodes are collapsed; internal nodes had good phylogenetic support with bootstrapping / posterior probability values >88/>0.93, respectively. For these analyses the alignment regions were expanded to a total of 346 amino acids. Mammalian CYP2E sequences were used as an outgroup.

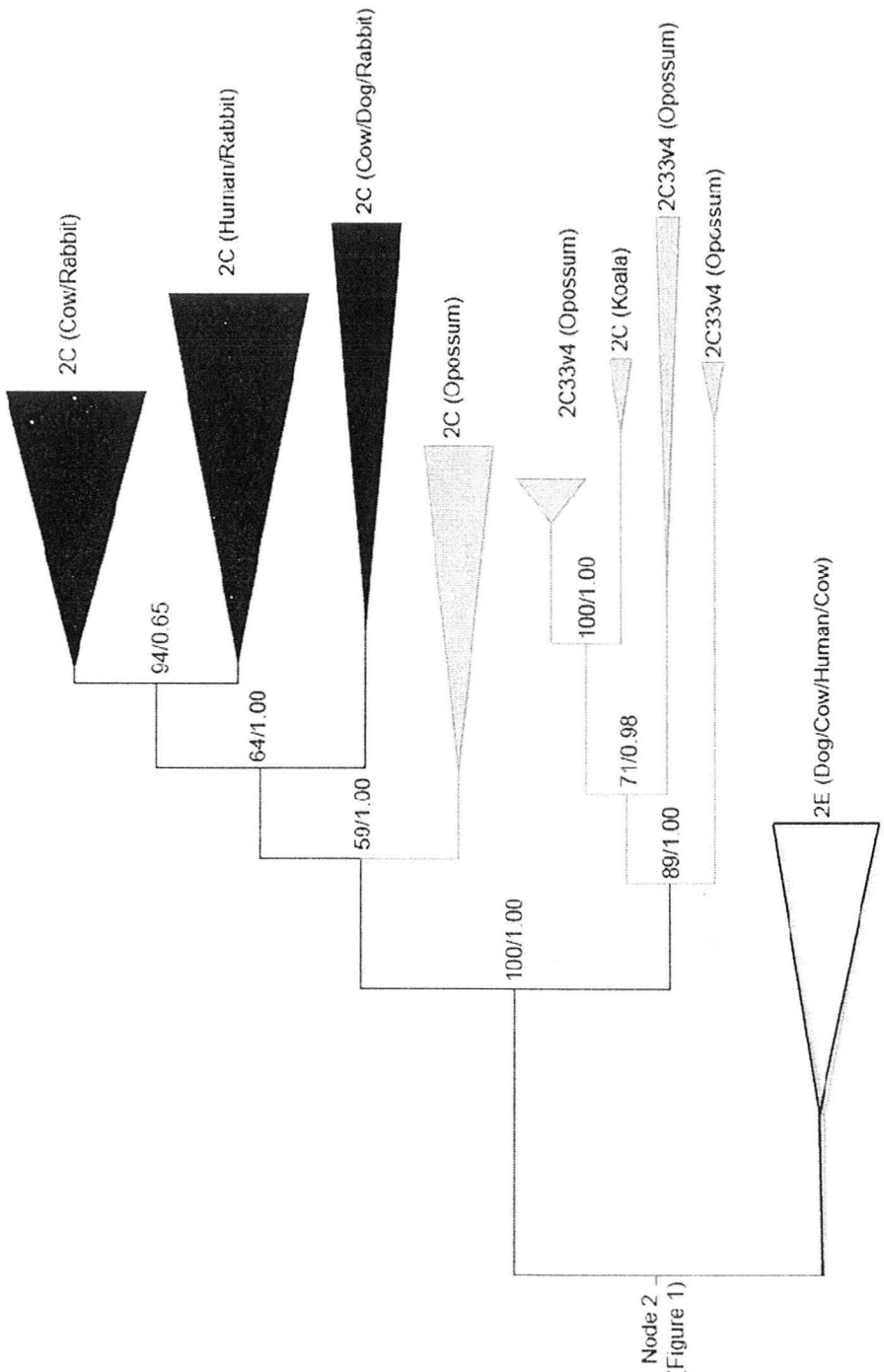

2C (Cow/Rabbit)
2C (Human/Rabbit)
2C (Cow/Dog/Rabbit)
2C (Opossum)
2C33v4 (Opossum)
2C (Koala)
2C33v4 (Opossum)
2C33v4 (Opossum)
2E (Dog/Cow/Human/Cow)
94/0.65
64/1.00
59/1.00
100/1.00
100/1.00
71/0.98
89/1.00
Node 2
(Figure 1)

**Figure 2.3. Phylogenetic analysis of the CYP2J-2Z cluster**. Phylogenetic analyses of the CYP2J-2Z cluster were conducted due to high complexity in the vertebrate CYP2 phylogeny (Node 6 and 7, Figure 1). A maximum likelihood tree is shown with bootstrapping and posterior probability values labeled on all distal nodes (See Materials and Methods for details). Mammalia and Amphibia, Actinopterygii, and outgroup (*Ciona*) clades are shaded black, light grey, and white, respectively. Internal nodes are collapsed; all internal nodes had good phylogenetic support with bootstrapping / posterior probability values >97/1.00, respectively. For these analyses the alignment regions were expanded to a total of 343 amino acids. Ciona CYP2N/2J and CYP2X sequences were used as outgroup clades.

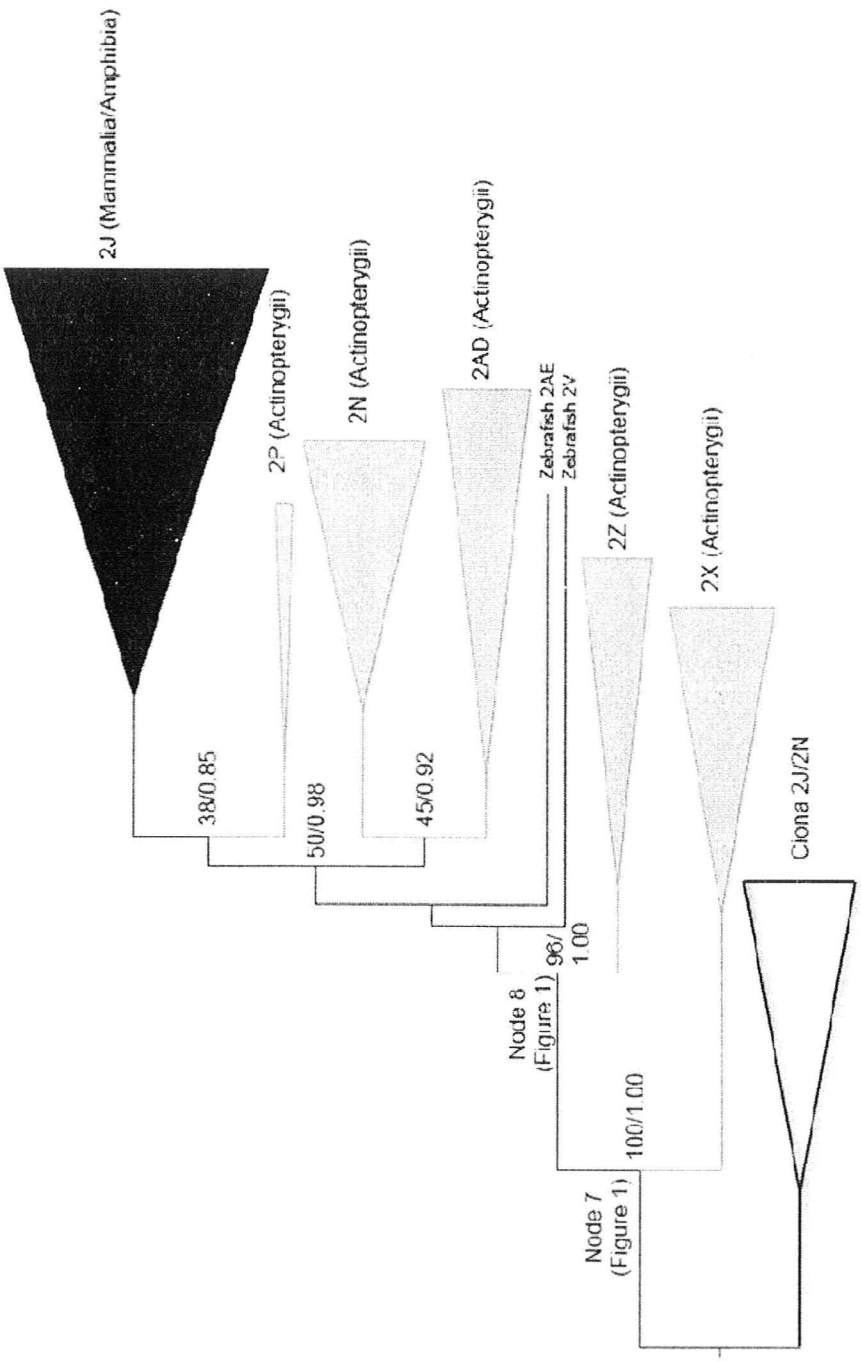

2J (Mammalia/Amphibia)
2P (Actinopterygii)
2N (Actinopterygii)
2AD (Actinopterygii)
Zebrafish 2AE
Zebrafish 2V
2Z (Actinopterygii)
2X (Actinopterygii)
Ciona 2J/2N
38/0.85
50/0.98
45/0.92
96/
1.00
Node 8
(Figure 1)
100/1.00
Node 7
(Figure 1)

**Figure 2.4. CYP2 Sequence schematic and DIVERGE heat map of 17 CYP2 subfamily pairwise comparisons.** Figures 6A and 6B are aligned to scale using the CYP2C8 sequence. **A)** A scaled schematic of the CYP2 gene that illustrates helices (A-L) and Substrate Recognition Site (SRS) sites (1-6). The heat map **(B)** reflects 136 CYP2 subfamily pairwise comparisons in a graphical representation of DIVERGE analyses. Masked regions not utilized for DIVERGE or phylogenetic analyses are shaded grey. Black regions indicate amino acid positions with a coefficient (theta= $\theta$) of evolutionary rate of functional divergence less than 0.5; in comparison, the red regions represent DIVERGE theta ($\theta$) $\geq$0.5. Mapped CYP2 helices (A-L) are labeled above the heat map; the right side of the rows indicates pairwise subfamily comparisons; start residue positions for unmasked regions are labeled below. In particular, the labeled CYP2A, CYP2E and CYP2N regions represent a total of 35/48 subfamily pairwise comparisons of CYP2A or CYP2E.

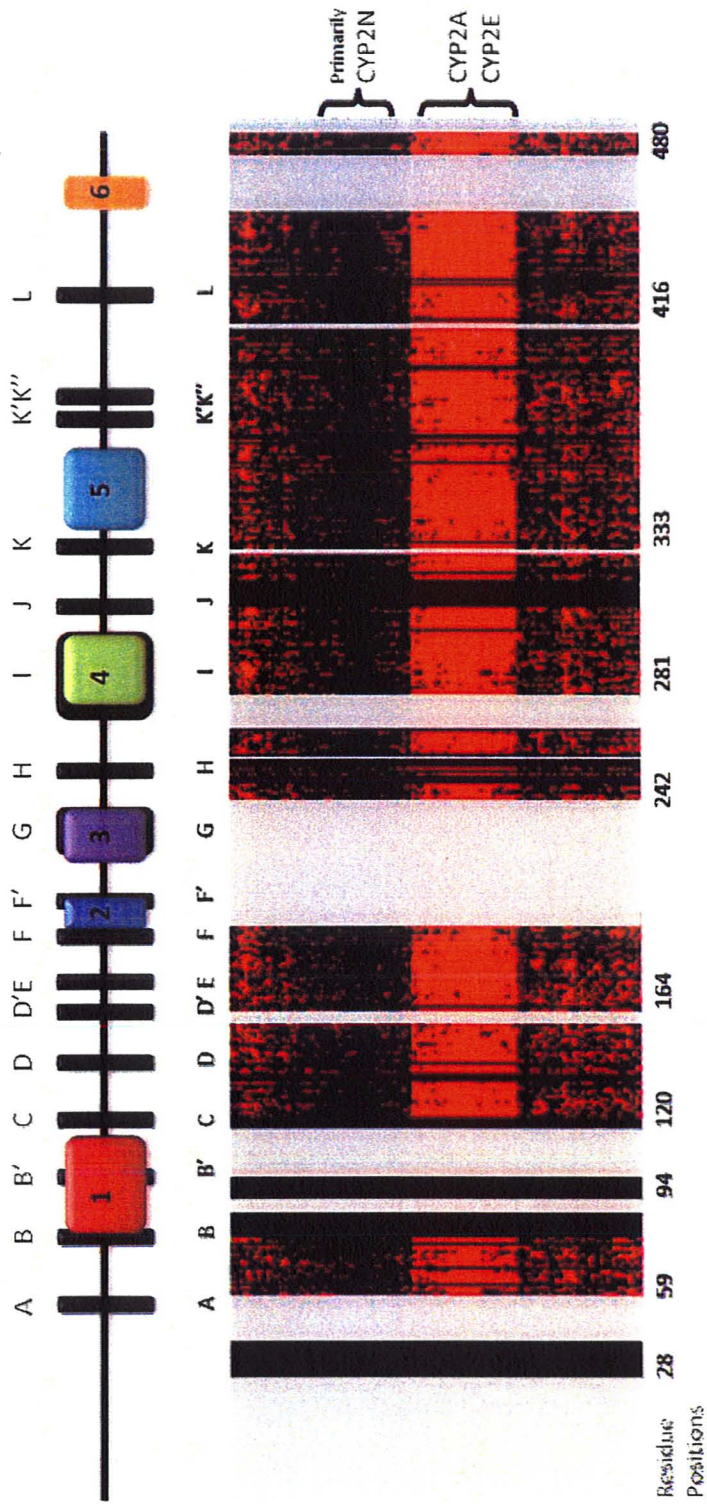

Primarily CYP2N
CYP2A CYP2E
A B B' C D D'E F F' G H I J K K'K'' L
1 2 3 4 5 6
A B B' C D D'E F F' G H I J K K'K'' L
28 59 94 120 164 242 281 333 416 480
Residue Positions

**Figure 2.5. Speciation patterns and gene duplications in the CYP 2U, 2R, and 2D subfamilies.** The evolution of vertebrate subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ). Dashed lines represent areas of topology with weak support. Four sequences (Amphibia 2R, 2U, and Aves 2U, 2D) were not detected within the present analysis; however, supportive data was identified and the sequence topology adapted (Thomas et al. 2007).

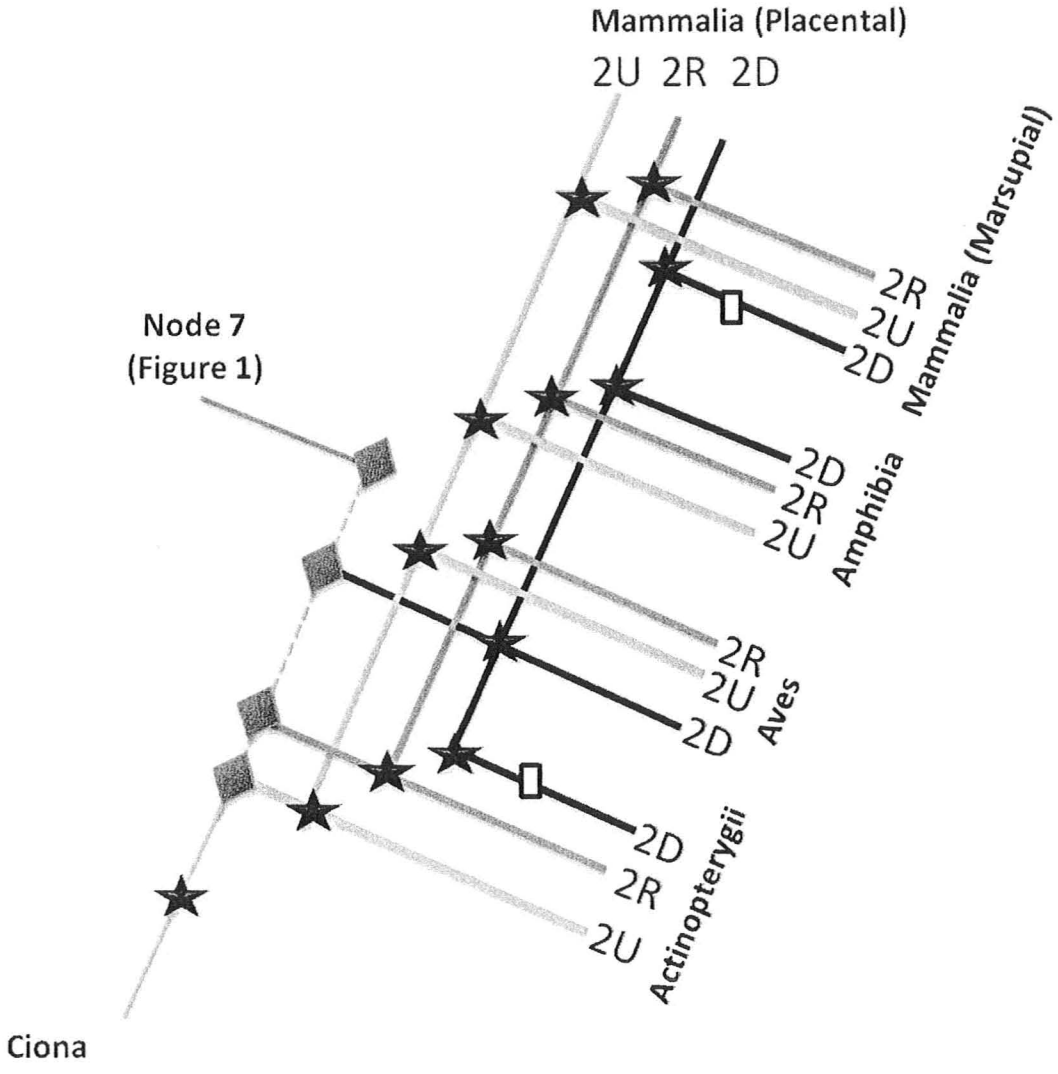

Mammalia (Placental)
2U 2R 2D
Mammalia (Marsupial)
2R
2U
2D
Node 7
(Figure 1)
2D
2R
2U
Amphibia
2R
2U
2D
Aves
2D
2R
2U
Actinopterygii
Ciona

**Figure 2.6. Speciation patterns and gene duplications in the CYP 2C, 2E, and 2H subfamilies.** The evolution of vertebrate subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ). Chicken 2C45 sequence is not a true 2C sequence, it does not cluster with the 2C sequences; and the sequence identity to other CYP2C sequences is less than 70%. Opossum CYP2 (XP_001364901) positioned between 2E (mammalian) and 2H (avian) subfamilies may be a putative CYP2E sequence due to its strong phylogenetic placement and sequence identify >70% to 2E subfamily sequences (Supplementary Figure 1).

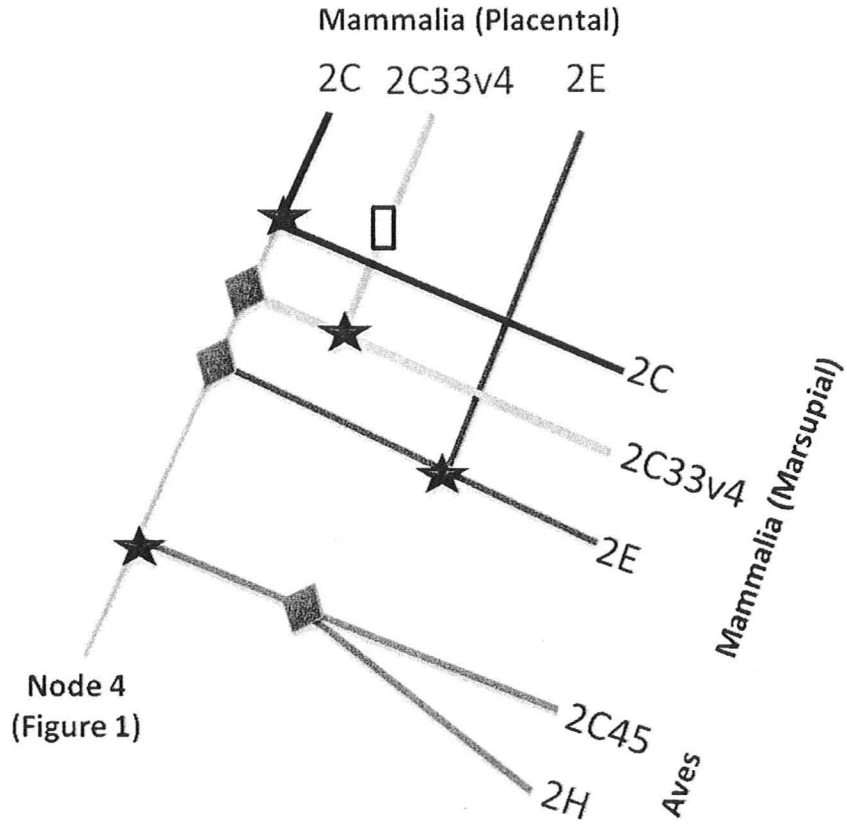
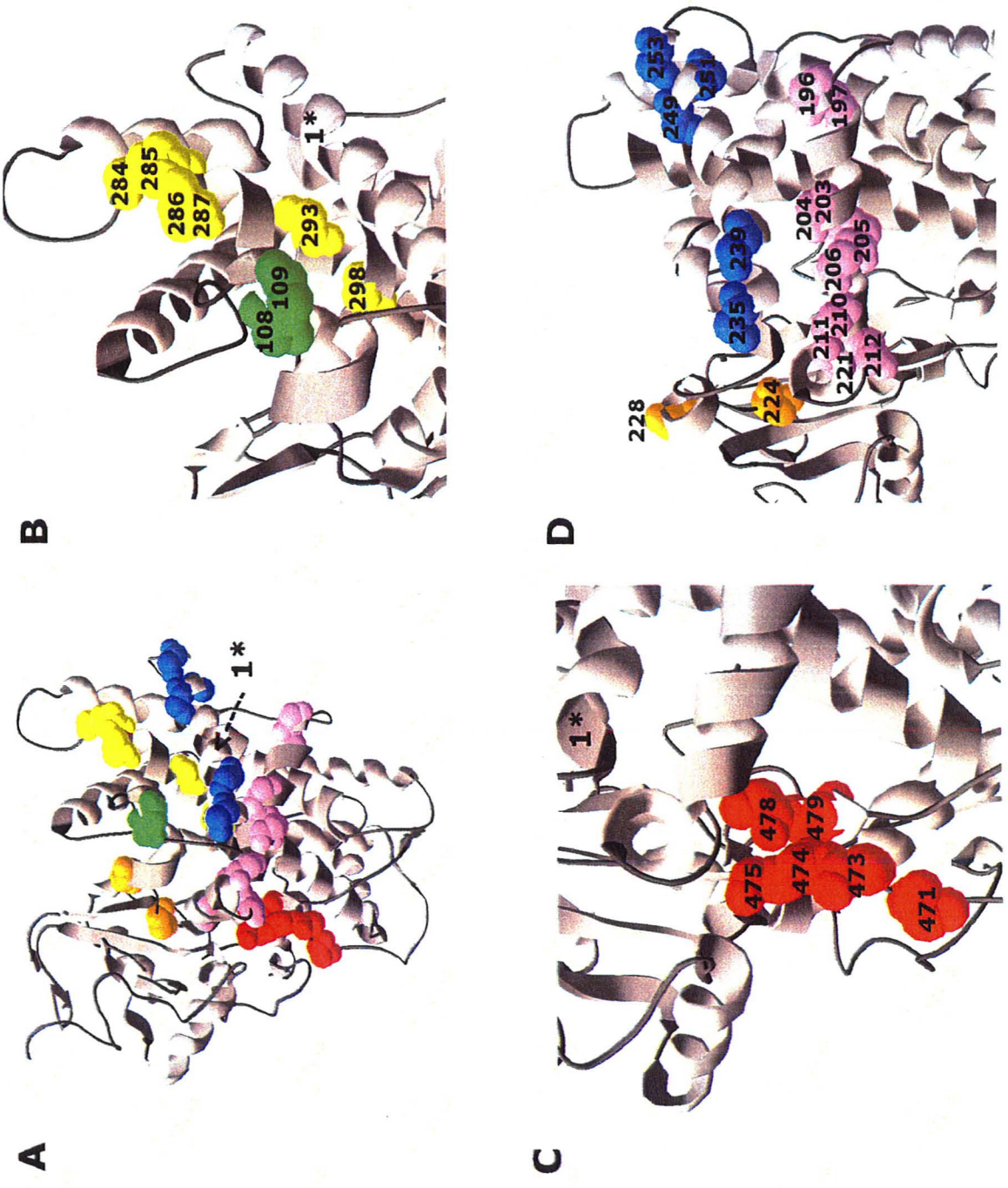

Mammalia (Placental)
2C
2C33v4
2E
2C
2C33v4
2E
Mammalia (Marsupial)
Node 4
(Figure 1)
2C45
2H
Aves

**Figure 2.7. Radical changes in the active site between CYP2F and CYP2A subfamilies.** Type II functional divergence of biochemical changes identified 31 radical changes ($\theta$>1) between the CYP2F and CYP2A subfamilies. These sites are shown in colour mapped on the hypothetical human CYP2F1 crystal structure (A). Helix G is identified in panels A, B, and C by 1* for orientation of the protein. The protein structure is rotated and enlarged to clearly show changes in the (B) N-terminus of helix I (yellow) and C-terminus of helix B (part of SRS1 green); (C) SRS6 (B4-1 and B4-2 sheets, red); (D) helix F (which includes SRS2, light purple), helix G (which includes SRS3, light blue) and the FG-loop (orange).


A
1*
B
284
285
286
287
293
1*
108
109
298
C
1*
478
475
479
474
473
471
D
253
249
251
196
197
204
203
239
205
235
206
211
210
221
212
228
224

# CHAPTER 3:

# PHYLOGENETIC AND FUNCTIONAL ANALYSES OF THE CYTOCHROME P450 FAMILY 4

Nina L. Kirischian and Joanna Y. Wilson

Department of Biology McMaster University Hamilton, Ontario, L8S 4K1, Canada

**Abstract**

The CYP4 family genes metabolize fatty acids, eicosanoids, and vitamin D and are important for chemical defense. The phylogenetic relationships among the CYP4 vertebrate genes are unclear and the function of CYP4 genes in non-mammalian vertebrates has not been empirically determined. In particular, existing phylogenies have not resolved the relationships between genes in the CYP4A, CYP4X, CYP4Z, CYP4B and CYP4T subfamilies. To better understand the evolutionary relationship of CYP4 subfamilies, 132 CYP4 sequences from 28 different species were utilized for phylogenetic reconstructions. Monophyletic groups were seen in most subfamilies; however monophyletic clades were lacking for the CYP4T and CYP4B subfamilies, suggesting that the nomenclature of these genes needs reconsideration. The phylogenetic reconstructions tracked known speciation patterns within CYP4 subfamilies, supporting the tree topology. CYP4V, the only known CYP4 subfamily in both vertebrates and invertebrates, was clustered with the invertebrate CYP4 subfamilies. The evolutionary rates of functional divergence were assessed between subfamilies in distinct taxonomic classes using type I divergence analyses. Evolutionary rates of functional divergence were high in pairwise comparison with CYP4X yet, pairwise comparisons with mammalian CYP4F22 genes generally had no statistically significant divergence. Type II functional divergence determined residues with radical biochemical changes between subfamilies and striking results were found in comparisons among the CYP4A, CYP4X, and CYP4B (Mammalia) subfamilies. Radical changes were detected in highly variable regions of the protein, which are suggested to be associated with substrate binding and

active site regions. Lastly, gene expression patterns were assessed *in silico* with EST libraries from common organs in human, chicken, frog and fish. CYP4V expression was markedly different between human (ubiquitous expression) and actinopterygian species (brain, kidney and liver organs); raising the hypothesis that CYP4V function has diverged across vertebrate species. The function of CYP4X, based on *in silico* functional divergence and gene expression data was hypothesized to specialize in metabolism of long chain fatty acids and may be important for neuromuscular tissue.

**Keywords**: cytochrome P450, CYP4 phylogenetic analysis, ω-hydroxilation, *in silico* functional divergence (type I and II), EST gene expression, CYP4F22

**Contributions:** This chapter has been formatted as a paper prepared for publication. Experimental work and data analysis was carried out by N.L.K. under the guidance and supervision of J.Y.W.

## 3.1 Introduction

The cytochrome P450 (CYP) 4 family is part of a superfamily of hemoproteins and is primarily responsible for ω-side fatty acid metabolism. The CYP4 family evolved approximately 1.25 billion years ago, following the development of steroid biosynthesis genes (Simpson 1997). *Mycobacterium leprae* CYP102, a bacterial CYP gene, clusters closely with the CYP4 clan and hydroxylates fatty acids, suggesting that fatty acid metabolism was one of the primitive functions of CYPs (Nelson 1998). The CYP4 family is divided into 72 subfamilies; the majority of subfamilies are in invertebrates (Baldwin *et al.* 2009; Nelson 2009; Tijet *et al.* 2001) and only 7 subfamilies are found in vertebrates (Hardwick 2008; Nelson 2003). The CYP4V subfamily is the only subfamily shared in both vertebrates and invertebrates (Nelson 2009). Nomenclature rules were not strictly enforced for the CYP4 family, as the CYP4V sequences are less than 40% identical to other vertebrate CYP4 sequences (Nelson *et al.* 2004). Yet, phylogenetic analyses of invertebrate CYP4 sequences support the inclusion of the CYP4V genes within the CYP4 family (Baldwin *et al.* 2009; Fujita *et al.* 2004; Nelson 2009; Nelson *et al.* 2004) suggesting that strict amino acid identity may not be sufficient to place some CYP genes within the CYP4 family, an approach that has been largely successful in other CYP families (Nelson *et al.* 2004).

The diversity of invertebrate CYP4 subfamilies is likely a strong indicator of functional diversity and adaptation to their diet (Baldwin *et al.* 2009; Davies *et al.* 2006; Feyereisen 2006) and phase-I detoxification system (Baldwin *et al.* 2009). A large

number of the insect CYP genes are found in the CYP4 family (Baldwin *et al.* 2009; Nelson 1998; Tijet *et al.* 2001), the genes are inducible, and they metabolize xenobiotics, alkaloids, nicotine and pheromones (Feyereisen 2006). The phylogenetic relationships among insect CYP4s has been studied in detail and there is a comprehensive understanding of the invertebrate CYP4 subfamily evolution (Baldwin *et al.* 2009; Dunkov *et al.* 1996; Feyereisen 2006; Rewitz *et al.* 2006; Srivastava *et al.* 2010; Tijet *et al.* 2001). The most comprehensive phylogenetic reconstruction of insect CYP4s illustrated monophyletic relationships of invertebrate CYP4 subfamilies with very strong phylogenetic support (Baldwin *et al.* 2009). Few evolutionary studies of the CYP4 family have included both invertebrates and vertebrates sequences (Baldwin *et al.* 2009; Nelson 2009; Tijet *et al.* 2001); yet the topologies are different, not all subfamilies were included and the number of sequences per subfamily were limited. The CYP4V subfamily is the only subfamily shared across metazoans and a few studies suggest that this subfamily is more highly related to invertebrate CYP4s than vertebrate CYP4s (Baldwin *et al.* 2009; Fujita *et al.* 2004; Nelson *et al.* 2004).

There are 7 CYP4 subfamilies in vertebrates (Nelson 2003; Nelson 2009). Of these subfamilies, several (CYP4X, CYP4A, CYP4Z) are specific to mammalian species (Hardwick 2008; Nelson 2003; Nelson *et al.* 2004; Savas *et al.* 2005); CYP4Z is only found in humans (Nelson *et al.* 2004; Rieger *et al.* 2004). In general, there is agreement for the placement of CYP4V and CYP4F subfamilies (Fujita *et al.* 2004; Nelson 2003; Nelson *et al.* 2004; Thomas 2007). The CYP4V subfamily clusters more closely with invertebrate CYP4 genes and the remaining vertebrate CYP4 subfamilies arose from a

common duplication event (Fujita *et al.* 2004; Nelson 2003; Nelson 2009; Thomas 2007). Interestingly, the CYP4A/4X/4Z/4B subfamilies cluster and were suggested to have evolved from the CYP4T subfamily, a family found in actinopterygian (Nelson 2003) and amphibian species (Fujita *et al.* 2004; Liu *et al.* 2009; Nelson 2003; Nelson 2009; Thomas 2007). While all existing studies cluster CYP4A, CYP4X, CYP4Z, CYP4B, and CYP4T together, the internal topology of this cluster differs across phylogenetic reconstructions (Fujita *et al.* 2004; Nelson 2003; Thomas 2007). Thus, the relationship among of the majority of the vertebrate CYP4 subfamilies is unclear.

Vertebrate CYP4 enzymes catalyze the $\omega$-hydroxylation of the terminal carbon of fatty acids, including essential signaling molecules such as eicosanoids, prostaglandins and leukotrienes (Hardwick 2008). Of the seven vertebrate subfamilies, three (CYP 4A, 4B, and 4F) have been studied in depth in mammals with respect to their function and gene expression (Hardwick 2008; Simpson 1997). More limited data is available for the function of the CYP4V (Fujita *et al.* 2004; Li *et al.* 2004), CYP4X (Al-Anizy *et al.* 2006; Savas *et al.* 2005; Stark *et al.* 2008), CYP4Z (Rieger *et al.* 2004; Savas *et al.* 2005), and CYP4T (Liu *et al.* 2009; Sabourault *et al.* 1998) subfamilies. The CYP4 enzymes metabolize fatty acid substrates of varied length, level of saturation and branching (Hardwick 2008). Specific subfamilies have shown preferences for the length of fatty acids; CYP4B, CYP4A and CYP4V, and CYP4F preferentially metabolize short (C7–C10), medium (C10–C16), and long- very long (C18–C26) fatty acids chains, respectively (Hardwick 2008). The metabolism, expression and function of mammalian

CYP4s have been reviewed in depth elsewhere (Hardwick 2008; Hsu *et al.* 2007; Simpson 1997).

In general, the CYP4B subfamily specializes in ω-hydroxylation of short chain fatty acids and metabolic exogenous compounds such as valporic acid ($C_8$), 3-methylindole ($C_9$) and several aromatic amines that are pro-toxic, that have potential to result in specific tissue toxicity (Baer and Rettie 2006). CYP4A enzymes are induced by starvation and an enriched fat diet (Hardwick 2008; Waxman 1999), however, functional and metabolic variability between different species exists (Hsu *et al.* 2007). Interestingly, metabolism of lauric and palmiric acid was detected with both CYP4A (Hoch *et al.* 2000; Kawashima *et al.* 2000) and CYP4V (Hardwick 2008; Nakano *et al.* 2009) proteins. CYP4As are known for catalyzing arachidonic acids to 20-HETE (Bellamine *et al.* 2003; Hardwick 2008). While the CYP4A and CYP4V subfamilies share some functional similarities, expression patterns differ. CYP4A was found in kidney and liver (Cowart *et al.* 2002; Palmer *et al.* 1993; Savas *et al.* 2005) whereas, CYP4V was expressed ubiquitously (Fujita *et al.* 2004; Li *et al.* 2004).

CYP4F subfamily enzymes are more diverse in the metabolic specificities yet the subfamily is known for ω-hydroxylation of very long fatty acids (VLFA; $C_{18}$-$C_{26}$), leukotrienes, prostaglandins, vitamins with long alkyl side chains, and HETE (Hsu *et al.* 2007; Jin *et al.* 1998; Kalsotra *et al.* 2004). The ω-hydroxylation of pro- and anti-inflammatory leukotrienes is a capability of CYP4F2 and 4F3, while CYP4F8 and 4F12 metabolize prostaglandins, endoperoxides and arachidonic acid (Bylund *et al.* 2000;

Hardwick 2008). The CYP4F11 and 4F12 genes can not only metabolize VLFA but are unique in the CYP4F subfamily since they hydroxylate xenobiotics such as benzphetamine, ethylmorphene, erythromycin (Kalsotra *et al.* 2004) and ebastine, (Hardwick 2008). The majority of the mammalian CYP4F genes were expressed in liver, intestine and kidney, yet the expression of certain CYP4Fs was not limited to these 3 organs (Cui *et al.* 2000; Kikuta *et al.* 2007).

Limited functional understanding exists for CYP4X and CYP4Z subfamilies. CYP4X gene expression was strongly associated with the brain (Choudhary *et al.* 2005; Hedlund *et al.* 2001) and the neurovascular regions. CYP4Z was primarily expressed in mammary glands and this expression was upregulated in breast cancer patients (Rieger *et al.* 2004; Savas *et al.* 2005). The existing data on mammalian CYP4 genes strongly suggest that the family is diverse in both physiological function and expression patterns.

CYP4 functional studies outside of mammals are limited, thus the understanding of CYP4 subfamilies in other vertebrate classes is primarily extrapolated from mammals. CYP genes from the same subfamily are often hypothesized to have similar function across vertebrate species. Gene expression data may be utilized to develop a better understanding of tissue specificity and derive hypotheses of putative gene function. We would expect similar patterns of expression in mammalian and non-mammalian species, should function be conserved within a subfamily. *In silico* techniques, using expressed sequence tags (ESTs), are a novel approach to raise functional hypotheses (Nagaraj *et al.* 2007). EST libraries are often available for species with completed genomes as part of

the annotation efforts and this data can be used to identify unique genes and their putative function based on their tissue expression (Nagaraj *et al.* 2007).

The evolution of CYP4 genes in metazoans has not been extensively studied and the phylogenetic relationships within vertebrate subfamilies are not clear. The present study includes CYP4 sequences from 4 major vertebrate (Mammalia, Aves, Amphibia and Actinopterygii) and several major invertebrate (Ascidiacea, Echinoidea, Gastropoda and Insecta) classes. We targeted species with complete genomes to ensure full coverage of CYP4 subfamilies. Genome annotation of CYP4 sequences (cow, dog, stickleback, medaka), existing sequences from previously annotated genomes (zebrafish, fugu, human) and exhaustive BLAST searching provided 132 CYP4 sequences in 28 different species for phylogenetic reconstructions. Type I functional analyses identified the evolutionary rates of functional divergence between 6 vertebrate subfamilies and a limited number of invertebrate clusters. Type II functional divergence determined sites with radical biochemical changes across vertebrate sister subfamilies only. Lastly, CYP4 gene expression levels and patterns were determined using EST data for human, chicken, frog, zebrafish, stickleback and medaka. This study provides a detailed understanding of the phylogeny of the CYP4 family in vertebrates and raises hypotheses regarding the function of the CYP4s across vertebrates.

## 3.2 Methods

### *3.2.1 CYP4 gene sequences*

A total of 132 CYP4 sequences were acquired via three methods: the Cytochrome P450 homepage BLAST server (Nelson 2009), *de novo* sequence prediction from genomic data, and through BLAST searching sequence databases (NCBI Genbank and Ensembl; Table 1; Supplementary Table 1). Mammalian and amphibian sequences were primarily retrieved from GenBank, actinopterygian and avian sequences were from Ensembl and the P450 homepage, respectively. The complete CYP4 gene complements had been previously identified in zebrafish (*Danio rerio*: Nelson 2009), fugu (*Takifugu rubripes*: Nelson 2003; Nelson 2009), and human (*Homo sapiens*: Nelson 2003; Nelson 2009; Nelson *et al.* 2004) and these sequences were retrieved from the Cytochrome P450 homepage BLAST server and NCBI. BLAST searching actinopterygian genomes with fugu CYP4 sequences produced BLAST hits with higher sequence identity than those with zebrafish sequences; fugu sequences were utilized to annotate CYP4 genes in the three-spined stickleback (*Gasterosteus aculeatus,* genome assembly version 1; V1) and medaka (*Oryzias latipes*, V1) genome assemblies. Human CYP4 sequences were the basis of annotation for CYP4 sequences in cow (*Bos taurus,* V4) and dog (*Canis familiaris*, V2.1) genomes. Insect sequences were represented by the full CYP4 complement from fruit fly (*Drosophila melanogaster*) retrieved from the P450 homepage. Lastly, the human and zebrafish CYP4 sequences were utilized in exhaustive and extensive BLAST searches in GenBank to identify additional CYP4 sequences in rabbit (*Oryctolagus cuniculus),* koala (*Phascolarctos cinereus),* opossum *(Monodelphis*

*domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis* and *X. laevis*), rainbow trout (*Oncorhynchus mykiss*), giant panda (*Ailuropoda melanoleuca*), goat (*Capra hircus*), horse (*Equus caballus*), minke whale (*Balaenoptera acutorostrata*), pig (*Sus scrofa*), sheep (*Ovis aries*), zebra finch (*Taeniopygia guttata*), European seabass (*Dicentrarchus labrax*), green pufferfish (*Tetraodon fluviatilis*), sea urchin (*Strongylocentrotus purpuratus*), sea squirt (*Ciona intestinalis*), and flamingo tongue snail (*Cyphoma gibbosum).* The BLAST searching strategy identified a significant number of CYP4 sequences for most vertebrate subfamilies. To root the CYP4 phylogenetic tree, sequences outside of the CYP4 family were chosen. Genes from the CYP46 family, a sister family to the CYP Clan 4 (Nelson 1999; Nelson *et al.* 2004) were selected as an appropriate outgroup. The evolutionary distance between the seven CYP46 sequences and the CYP4 sequences was quite large; 4 CYP4 sequences from fungi and choanoflagellates species were utilized to reduce the distance between clades in the base of the phylogeny.

The annotation approach for CYP4 sequences was the same for all *de novo* sequences. Existing sequences (zebrafish, fugu and human) were utilized in BLAST searches and genome regions with high identity (>60%) were retrieved. Mammalian CYP4A, CYP4B, CYP4X, CYP4Z and 4F sequences had 12 exons (Nelson *et al.* 2004) and CYP4V sequences had 11 exons (Nelson 2009). The actinopterygian CYP4V, CYP4T and CYP4F sequences had 11, 12, and 13 exons, respectively, based on zebrafish CYP4 genes (Nelson 2009). All vertebrate CYP coding transcripts are approximately 1500 base pairs in length. BLAST searches with high sequence

similarity and 11-13 exons found in appropriate order (i.e. exon 1 must be adjacent to exon 2, etc.) were retrieved. Putative nucleotide sequences were imported into MacClade 4.0 (Maddison and Maddison 2000) for annotation of exon and intron boundaries, based on derived eukaryotic consensus splice sites (Rogers and Wall 1980). Putative CYP4 sequences were assessed for accurate splice site boundaries via BLAST 2 Sequence server (i.e. dog CYP4V vs. human CYP4V). For quality assurance, all genes were assessed for the appropriate number of exons (11-13), correct exon order, a total length of ~1500 bp for the coding transcript and the presence of start and stop codons. For *de novo* sequences, RefSeq and EST data were used, where possible, to support the annotated sequences. All *de novo* annotated sequences were submitted to the P450 nomenclature committee for sequence verification and nomenclature assignments.

*3.2.2 CYP4 gene alignment*

The CYP4 amino acid sequences from 28 species were aligned with CYP46 sequences using CLUSTALX (Thompson *et al.* 1997), and the procedures were similar to those in Kirischian *et al.* (2011). Manual adjustments to the alignment were performed using Mesquite 2.0 (Maddison and Maddison 2000). Alignment regions with gaps and uncertain homology were excluded (masked) from the final phylogenetic analyses. The final analyses were based on 143 aligned sequences with a total length of 331 amino acids.

Identifying protein domains (helices, beta sheets) within the CYP4 alignment was difficult when comparing CYP4 sequences to the CYP sequences in Gotoh (1992).

To increase the confidence in identification of the helix regions, the human CYP4B1 sequence was superimposed onto the known 3D crystal structure of CYP2D6 (Rowland *et al.* 2006), using Swiss-Pdb Viewer Deep view software (Guex and Peitsch 1997). Based on the hypothetical crystal structure of CYP4B1 and alignments in several studies (Gotoh 1992; Lewis 2002), putative protein domains were mapped onto the CYP4 alignment.

### *3.2.3 Phylogenetic analyses*

The phylogenetic reconstructions of CYP4 evolutionary history were determined using maximum likelihood and Bayesian inference approaches. The maximum likelihood phylogeny was based on the WAG+I+G (Ren *et al.* 2005; Whelan and Goldman 2001) substitution model with unequal amino acid frequencies, as determined by ProtTest (Abascal *et al.* 2005). The maximum likelihood analyses were computed on a RAxML (Randomized Axelerated Maximum Likelihood) BlackBox server (Stamatakis *et al.* 2008) with 1000 bootstrapping replicates. Bootstrapping analysis was computed using maximum likelihood with an estimated proportion of invariant sites and empirical base frequencies with the indicated outgroup (CYP46) sequences.

Phylogenetic reconstructions with Bayesian inference were completed using the MrBayes computer software program (V3.1.1: Huelsenbeck *et al.* 2001). The $MC^3$ (Metropolis-coupled, Markov chain, Monte Carlo) searches were performed using four incrementally heated chains with distinct random initial trees for 20 million generations,

with a sampling frequency of 1000 generations. Posterior probabilities were estimated after the removal of $MC^3$ burn-in. Posterior probabilities for nodes that appeared in the maximum-likelihood tree but were not present in the consensus Bayesian phylogenetic reconstruction were assessed using 19, 970,000 sampled tree topologies from the Bayesian analysis in PAUP 4.0 (Swofford, 2000).

### *3.2.4 Functional divergence*

To develop a better understanding of CYP4 functional evolution, an analysis of the amino acid alignment in context of the maximum likelihood tree was conducted using DIVERGE (2.0; Gu and Vander Velden 2002). DIVERGE utilizes a phylogenetic tree to assess site-specific changes in evolutionary rates within amino acid alignments when comparing subclades (CYP4 subfamilies in our case). DIVERGE uses the coefficient of evolutionary functional divergence ($\theta$) to measure changes in site-specific evolutionary rates; $\theta = 0$ indicates no functional divergence while increasing values indicate increasing functional divergence, with $\theta = 1$ being the maximum. Utilizing the coefficient of evolutionary functional divergence ($\theta$), we tested for significant functional divergence (likelihood ratio test (LRT), $p<0.05$) for each of the pairwise comparisons of 13 different CYP4 subfamily groups, which were subdivided into taxonomic groups where appropriate, based on the phylogenetic tree. For example, the CYP4F sequences were divided into mammalian CYP4F22, placental mammalian CYP4F (except CYP4F22 genes), and actinopterygian CYP4F sequences;

these three groups represent monophyletic clades in the phylogenetic trees with greater than 4 sequences per clade, as required by DIVERGE.

Using DIVERGE type II functional analyses, site-specific posterior probabilities identified amino acids with radical biochemical changes between sister subfamily groups. We focused on radical changes within the active site and substrate binding domains and applied a cut-off value of $\theta > 1$ for site-specific posterior probabilities, an approach used in Kirischian *et al.* (2011). Type II functional analyses were completed for sister subfamilies; groups were divided into vertebrate classes, when appropriate, for these analyses. Type II functional divergence was analyzed for the following groups: mammalian CYP 4A/4X/4B; CYP4B (mammalian)/4T (actinoperygian)/4T (amphibian); CYP4F (mammalian)/4F (actinopterygian)/4F22 (mammalian), 4F22 (mammalian)/4T (actinopterygii); and CYP4V (mammalian)/4V (actinopterygian). The masked regions of the sequence alignment were reassessed for each group, and since there was higher similarity in sister subfamilies than in the full CYP4 alignment, a larger number of residues were included in the type II functional divergence.

### *3.2.5 Digital gene expression analysis*

Expressed sequence tag (EST) data were retrieved from Unigene to determine tissue specificity and intensity of expression of CYP4 genes in common organs across species in different vertebrate classes. The EST libraries were accessed for each CYP4 gene of interest at dbEST via BLAST on NCBI (Boguski *et al.* 1993) and common ESTs were clustered into non-normalized EST profiles in UniGene (NCBI 2010). EST data

was normalized for the size of the EST library and expressed in transcripts per million (TPM). Libraries range from 1 000 to over a million entries; data was used from libraries greater than 10,000 sequences to ensure reasonable data quality (Schmitt *et al.* 1999). No EST library above 10,000 ESTs was available for fugu. The TPM data had an exponential distribution and was log transformed to generate a normal distribution in statistical analysis software (SAS); using a mean value of 32 TPM and the lower interval of 95% confidence around the mean, we applied a cut-off value of 6 TPM as the minimum expression considered significantly different from 0. TPM data was collected from at least one species for each of the four vertebrate classes mammalian (human), amphibian (frog), avian (chicken), and actinopterygian (zebrafish, stickleback and medaka). Significantly more EST data was available for zebrafish than either stickleback or medaka. Expression patterns were examined for four subfamilies or subfamily clusters, based on their position in the phylogenetic tree: CYP4V, CYP4B and CYP4T, CYP4A, CYP4X and CYP4Z, and CYP4F. Data from four common organs (brain, kidney, liver, and lung) were found in all vertebrate species for each EST subfamily comparison.

## 3.3 Results

### *3.3.1 CYP4 annotation and nomenclature*

The CYP4 analyses included 132 CYP4 genes representing 17 CYP4 subfamilies; 23 sequences were *de novo* annotations (Table 1). Accession numbers of each sequence are provided in Supplementary Table 1. The CYP4 genes within one

subfamily were consistent in exon number for a given vertebrate class; yet, exon numbers were sometimes different within the same subfamily in different vertebrate classes. Genome annotations for actinopterygian species (stickleback and medaka genomes) identified 11, 12, and 13 exons for CYP4V, CYP4T, and CYP4F genes, respectively (data not shown). Based on those species with completed genomes, the total number of CYP4 genes per species varied with vertebrate class; actinopterygian, amphibian, avian, and mammalian species had 3-4, 5-6, 2-3, and 8-12 sequences, respectively (Table 1).

*3.3.2 CYP4 phylogenetic analyses*

The CYP4 phylogeny was reconstructed with both maximum likelihood and Bayesian inference approaches. In the Bayesian inference, $MC^3$ burn-in length was 30, 000 of 20,000,000 generations, resulting in 19,970,000 sampled trees for the calculation of posterior probabilities. The overall topology of the maximum likelihood tree (Figure 1) was in agreement with the Bayesian inference consensus tree; the maximum likelihood tree, with all internal nodes, is provided in Supplementary Figure 1. Internal nodes in the CYP4 phylogeny mostly had strong bootstrapping (>75%) and posterior probability (>0.80) values (Supplementary Figure 1). Most vertebrate subfamilies clustered monophyletically, with patterns of internal branching following known vertebrate (Li *et al.* 2007; Prasad *et al.* 2008) and invertebrate speciation patterns with some exceptions. Surprisingly, chicken_CYP4A/4B was strongly placed with a zebra finch CYP4B gene instead of with the mammalian CYP4A genes but the support for

placement of the avian CYP4A/B and other 4B genes within the tree was not strong (Supplementary Figure 1).

Several sequences were lacking nomenclature and for most there was strong phylogenetic support for subfamily placement. Based on our phylogeny, subfamily based nomenclature can be suggested for: sea urchin_ CYP4V_ XP_783176, sea urchin _CYP4V_ XP_783244, sea urchin _CYP4V_ XP_784930, sea urchin _CYP4V_ XP_786946, snail_CYP4V_ ACD75824, green pufferfish_ CYP4V_ CAF96593, giant panda_CYP4V_ EFB25443, giant_panda_CYP4F22_ EFB18761, opossum_ CYP4F_ XP_001367797, and pig_CYP4A_ NP_999590. New subfamily assignments should be considered for *Ciona* and snail CYP4F sequences (*Ciona*_CYP4F17_XP_002130606, *Ciona*_CYP4F22_XP_002129693, *Ciona*_CYP4F40_XP_002132033, snail CYP4F_ ACD75827, snail CYP4F_ ACD75830, and snail CYP4F_ ACD75825) since they did not cluster within the CYP4F subfamily.

### *3.3.3 CYP4 phylogenetic analyses*

There are two major clades of CYP4 subfamilies on the phylogenetic tree and this bifurcation is strongly supported by bootstrapping (98%) and posterior probabilities (1.00; Node A, Figure 1). First, 10 invertebrate CYP4 subfamilies (CYP 4C, 4D, 4E, 4G, 4P, 4S, 4AA, 4AE, 4AD and 4AC) and all CYP4V genes clustered with strong support (Node B, Figure 1). All 22 *Drosophila* sequences clustered together in a distinct clade from the CYP4V genes, although with poor internal node resolution for the insecta CYP4 subfamilies (Supplementary Figure 1). The CYP4V subfamily clade included sequences

from both vertebrate and invertebrate species, including one snail and *Ciona* sequence and 4 sea urchin sequences. The placement of the *Ciona* CYP4V and sea urchin CYP4 sequences was unresolved but there was strong support for the placement of the snail CYP4V sequence at the base of the vertebrate CYP4V sequences. The topology of the vertebrate CYP4V sequences (Node C, Figure 1) followed known patterns of vertebrate evolution.

The second major clade (Node D, Figure 1) in the CYP4 phylogenetic reconstruction includes all subfamilies found in vertebrates, except CYP4V. At the base of this clade are 3 snail CYP4 sequences and 3 *Ciona* CYP4F sequences, although the support was only strong for the placement of the snail sequences. The CYP4F subfamily was a monophyletic clade of the vertebrate sequences; the previously identified *Ciona* CYP4F sequences did not cluster with the vertebrate CYP4F sequences. The mammalian CYP4F22 clade was strongly supported as the root of the vertebrate CYP4F clade (Node E, Figure 1); the other CYP4F sequences were grouped according to their taxonomic class. Although there was strong support for the bifurcation of CYP4F and CYP4F22 genes (Node E, Figure 1), and for the clustering of CYP4F genes within each taxonomic class, the internal topology of the CYP4F clade (Node F, Figure 1) was poorly supported.

The vertebrate CYP 4T/4B/4X/4Z/4A cluster arose from a bifurcation event (Node G, Figure 1), separating the CYP4F genes from the majority of vertebrate CYP4 subfamilies. This cluster is rooted with strong support by the CYP4T (actinopterygian) clade (Node H, Figure 1). Surprisingly, the frog CYP4Ts did not cluster with those from

fish, nor did the avian CYP4Bs cluster with the mammalian 4B genes, although support for the placement of the avian CYP4B sequences was weak (Supplementary Figure 1). The CYP4B mammalian clade was placed at the root of the CYP4A, CYP4X and CYP4Z mammalian subfamilies; these subfamilies appear to share a common gene duplication event (Node I, Figure 1). CYP4X and 4Z were strongly supported sister subfamilies (Node J, Figure 1).

### *3.3.4 Type I functional divergence analyses*

Type I functional divergence was conducted on 6 vertebrate CYP4 subfamilies with the exception of the CYP4Z subfamily, which is composed of 1 human sequence. Using the masked alignment, the maximum likelihood tree topology, and the CYP46A1 crystal structure we determined the evolutionary rates of functional divergence of CYP4s using DIVERGE (Version 2.0). The coefficient of evolutionary functional divergence ($\theta$), its standard error, and the maximum likelihood ratio (LRT) were determined for each pairwise comparison (Supplementary Table 2). Considering the number of sequences per subfamily was quite large, sequences were grouped by both subfamily and taxonomic class, where possible, and were referred to as subfamily groups. Avian and amphibian sequences, with the exception of amphibian CYP4T, were excluded because DIVERGE requires a minimum of 4 sequences per group. A total of 13 groups were identified with sufficient sequences for this analysis. Mammalian groups were included from CYP4A, CYP4X, CYP4B, CYP4V, CYP4F and CYP4F22 subfamilies. Actinopterygian groups included CYP4T, CYP4V and CYP4F subfamilies. The remaining groups included the

amphibian CYP4T, sea urchin putative CYP4V, *Drosophila* CYP4D, and the fungi & choanoflagellates putative CYP4 genes. These 13 groups resulted in 76 pairwise group comparisons, of which 58 pairwise comparisons exhibited statistically significant divergence (LRT, $p<0.05$; Supplementary Table 2). A heat map of the site-specific evolutionary rates of functional divergence (Figure 2) identified a cluster of 22 pairwise comparison groups with highly divergent residues; pairwise comparisons with CYP4X dominated this cluster (Figure 2). The least divergence was found in comparisons with the mammalian CYP4F22 group. There were 6 distinct regions (helices B, B', J', K', K'', and a region on the C-terminus) within the CYP4 alignment that had low or no divergence (site-specific posterior probability, $\theta<0.5$) throughout all pairwise comparisons. In all cases, there was no evident clustering of divergent residues in any specific domain (Figure 2).

### *3.3.5 Type II functional divergence analyses*

Amino acids with radical biochemical changes between CYP4 subfamilies were identified via type II functional analyses. We did not perform all pairwise comparisons but performed pairwise comparisons within subfamily clades identified in the phylogenetic tree; specifically subfamilies CYP4V, CYP4F, and CYP4A/X/Z. CYP4T and CYP4B were also subjected to type II functional divergence analyses although these do not represent a single monophyletic clade on the phylogenetic tree. Statistically significant posterior probabilities for comparisons ranged from $\theta>1$ to 8 (data not shown). All comparisons had at least one site identified with radical change. For the CYP4V

subfamily, pairwise comparisons were between mammalian, actinopterygian, and sea urchin sequences. Comparisons between mammalian and actinopterygian CYP4V genes identified 17 residues with radical changes from a 389 amino acid alignment (data not shown). The site-specific posterior probabilities ranged from $1<\theta<4$ and the sites with radical biochemical changes were distributed throughout the alignment without any evident clustering. Few sites with radical biochemical changes were detected between the sea urchin and mammalian ($\theta$ =1.45 for 3 radical changes) or actinopterygian ($\theta$ =1.28 for 7 radical changes) CYP4V sequences.

Type II functional analyses in the CYP4F clade (Node E, Figure 1) were between sequences in actinopterygian and mammalian species; sequences from the mammals were divided into those from the CYP4F22 clade and the remaining CYP4F (CYP4F*) sequences. There were 19, 9 and 10 sites with radical changes in the comparisons between CYP4F*/CYP4F fish, CYP4F22/CYP4F*, and CYP4F fish/CYP4F22, respectively. There were no common sites identified between the three comparisons, however a couple of radical changes were identified in at least two of these analyses . In general, the radical changes were scattered throughout the protein with a focus from helix G to the C-terminus.

Pairwise comparisons between mammalian CYP4B, actinopterygian CYP4T, and amphibian 4T were completed for type II functional analyses. The largest number of radical changes was detected between CYP4T in actinopterygian and amphibian species, with 13 radical changes from a 418 amino acid alignment (data not shown). Only 2 and 4

changes were detected, in comparisons between mammalian 4B and either actinopterygian 4T or amphibian 4T, respectively. No clustering of sites with radical changes was evident.

The largest number or radical changes detected in type II functional divergence analyses were in comparisons between the mammalian CYP4A, CYP4X and CYP4B subfamilies. The sequences from CYP4B were only mammalian because the avian CYP4B sequences did not cluster monophyletically with the mammalian CYP4B sequences (Figure 1). A total of 79, 49 and 17 sites with radical changes were identified between CYP4A/4X, CYP4X/4B, and CYP4B/4A, respectively. The highest site-specific posterior probabilities were identified in the CYP4A/4X comparison and those found in the active sites of the protein are illustrated in Figure 3. A total of 29 sites with radical changes were shared between CYP4A/4X and CYP4X/4B analyses. Clusters of 3-9 sites with radical biochemical changes were detected from the CYP4A/4X comparison in helix F (3 sites, Figure 3B), F-G loop (3 sites, Figure 3B); helix G (7 sites, Figure 3B), helix I (6 sites, Figure 3C); between helices G and H (9 sites, Figure 3D); and helix J' (3 sites, Figure 3D). All three pairwise comparisons identified 3 sites with radical changes (residues 331, 335, 360 in the human CYP4X1 gene), these sites were located between helix I and J' (Figure 3C and 3D).

*3.3.5 Digital gene expression*

Quantitative analysis of expressed sequence tags (EST) was used to determine gene expression patterns in various tissue libraries for different vertebrate CYP4

subfamilies. Expression was measured in transcripts per million (TPM) for genes in the CYP4A/4X/4Z, 4F (Figure 4), 4B/4T, and 4V subfamilies (Supplementary Figure 2). Expression patterns for these genes were shown for 6 organs; 4 common organs were used in all analyses (brain, kidney, liver, and lung/gills). Human CYP4A11 and CYP4A22 gene expression (Figure 4A) was only found in two organs; both were found in liver and CYP4A11 was also found in kidney. Liver expression of CYP4A11 was three times higher than CYP4A22 (Figure 4A). Expression of CYP4Z and CYP4X genes were detected in 5 and 20 tissues, respectively. CYP4X1 expression was highest in trachea (324 TPM; Figure 4A), a level of expression that was significantly higher than that seen with any other CYP4A or CYP4Z gene. CYP4X1 strong expression was found in multiple tissue libraries including trachea, nerves, vascular, uterus, larynx liver, kidney, lung and mammary glands. CYP4Z1 gene expression was in mammary gland (97 TPM), trachea, liver (Figure 4A), testis (6 TPM) and connective tissue (100 TPM; data not shown. Liver was the only commonly shared organ that expressed CYP4A, CYP4X and CYP4Z genes.

The CYP4F family has five sequences in human and all but CYP4F22 are co-expressed in liver, kidney and intestine (Figure 4B). Zebrafish CYP4F43 gene expression was similar to that seen for the CYP4F11 and CYP4F3 human genes, except that the zebrafish gene was not identified in the liver EST library, yet, microarray data supports the presence of CYP4F43 in liver (Wilson J.Y., unpublished data). The amphibian CYP4F22 gene was expressed in liver and had high expression in intestine

(187 TPM). Interestingly, CYP4F40 was not expressed in the medaka liver tissue library (data not shown).

The human CYP4V2 gene was ubiquitously expressed (found in 28 out of 45 libraries); something that was not found in other vertebrate species (Supplementary Figure 2A). CYP4V genes were expressed, ranging from 50-179 TPM, in respiratory organs (lung/gill) and liver for all vertebrate species, with the exception of amphibian liver (CYP4V2). There was no EST library for lung in chicken (Supplementary Figure 2A). Exceptionally high levels of expression were in human parathyroid glands and ear for CYP4V2 (data not shown). Gene expression data for medaka and stickleback was very limited in both the number and size of libraries; libraries of sufficient size were available for liver in medaka and lung and brain in stickleback. CYP4Vgenes were expressed in the medaka liver (272 TPM) and stickleback gill (349 TPM); there was no expression of CYP4V genes in stickleback brain (data not shown).

At least one CYP4B and CYP4T gene was identified in the brain libraries from most vertebrate species with the exception of human (Supplementary Figure 2B). All CYP4B and CYP4T genes were expressed in intestine; there was no intestine EST library available for zebrafish (Supplementary Figure 2B). All genes, except human CYP4B1 and frog CYP4T9, were expressed in liver. Interestingly, only the human CYP4B1 gene was expressed in the lung; although there was no avian EST lung library (Supplementary Figure 2B).

## 3.4 Discussion

### *3.4.1 CYP4 nomenclature*

The CYP4 phylogenetic analysis was reconstructed using 143 sequences of which 132 were CYP4 genes (Supplementary Table 1; Supplementary Figure 1). The *de novo* annotated sequences were submitted to the P450 committee for quality verification and name assignments. The number of CYP4 genes per species varied; higher numbers were in mammalian genomes (8-12) and the lowest number in actinopterygian species (3-4), a similar difference in genes per species was seen in prior studies (Nelson 2009; Nelson *et al.* 2004). Yet, no extreme gene expansion was identified within a single CYP4 subfamily as has been noted in certain CYP2 subfamilies (Kirischian *et al.* 2011; Nelson *et al.* 2004); the only gene duplications were in CYP4A and CYP4F subfamilies (Figure 1; Supplementary Figure 1; Nelson *et al.* 2004). The mammalian CYP4F gene amplification has been associated with increased diversity in function in metabolizing both endogenous and exogenous compounds (Cui *et al.* 2001; Hardwick 2008; Kalsotra *et al.* 2004). The CYP4F11 enzyme metabolizes eicosanoids and drugs (i.e. erythromycin; Kalsotra *et al.* 2004); whereas, CYP4F2 and CYP4F3 ω-hydroxylate pro- and anti-inflammatory leukotrienes (Bylund *et al.* 2000; Hardwick 2008). Significantly greater expansion of CYP4 genes has been found in rodents in the CYP4A, CYP4B, and CYP4F subfamilies (Nelson *et al.* 2004); rodents were not included in this study.

This CYP4 phylogenetic reconstruction identified possible problematic nomenclature for several CYP4 genes. The assigned CYP4F subfamily to 3 *Ciona*

sequences (*Ciona*_CYP4F17_XP_002130606, *Ciona*_CYP4F22 XP_002129693, *Ciona*_CYP4F40 XP_002132033) may be incorrect, due to their basal placement outside of the CYP4F subfamily in the phylogenetic tree (Supplementary Figure 1; Figure 1); a new nomenclature assignment should be considered for these sequences. Amphibian sequences assigned to the CYP4T subfamily in our study have been identified as CYP4B sequences in Ensembl, NCBI and at least one published study (Liu *et al.* 2009). We have followed the nomenclature assigned by the P450 nomenclature committee and published by Nelson (2009). Interestingly, these genes and the avian CYP4B genes do not cluster in a single monophyletic clade with either actinopterygian CYP4T or mammalian CYP4B genes (Supplementary Figure 1), suggesting that the evolutionary history of these subfamilies will require more study when additional sequences become available.

### *3.4.2 Global overview of cytochrome P450 4 family phylogeny*

The CYP4 phylogenetic analyses resulted in a high resolution tree with strong bootstrap values (>80) and posterior probabilities (>0.95) for most distal and internal nodes (Supplementary Figure 1); the posterior probabilities are usually higher than bootstrap values (Whittingham *et al.* 2002). Monophyly was detected in all vertebrate subfamilies with the exception of CYP4T and CYP4B subfamilies and in all insect subfamilies. Only *Drosophila* sequences were included which provided too poor a taxonomic sampling for good internal node resolution in this portion of the phylogenetic tree thus internal nodes were not always strongly supported in this portion of the tree (Supplementary Figure 1). Other phylogenetic studies have provided a more detailed

and comprehensive insecta specific CYP4 reconstructions (Baldwin *et al.* 2009; Dunkov *et al.* 1996; Feyereisen 1999; Tijet *et al.* 2001).

Overall, tracked speciation patterns matched known vertebrate (Prasad *et al.* 2008) and invertebrate patterns (Adoutte *et al.* 1999; Adoutte *et al.* 2000; Ruiz-Trillo *et al.* 2008). For genes from the Actinopterygii class, CYP4 genes followed expected speciation patterns with zebrafish as most distal species, medaka as an intermediate and fugu and stickleback as recently evolved sister species (Li *et al.* 2007). For the CYP4 phylogenetic reconstruction, sequences outside of the CYP4 family were required to root the tree. Previous studies (Nelson 1999; Nelson *et al.* 2004) showed the CYP46 family to cluster closely with the CYP4 family, thus CYP46 genes were selected as the root in the CYP4 phylogenetic reconstruction. Since the evolutionary distance between CYP46 sequences and insect CYP4 sequences was large, sequences from distantly related organisms to animals such as putative CYP4 sequences from fungi and choanoflagellates were used to reduce the chance long-branch attraction (Bergsten *et al.* 2005) in the phylogenetic reconstruction.

### *3.4.3 Evolution of invertebrate CYP4 subfamilies and CYP4V*

The CYP4 phylogenetic tree has two major clusters; CYP4V sequences clustered with the invertebrate CYP4 subfamilies, while all other vertebrate CYP4 subfamilies group together (Node A, Figure 1). The diversification of CYP4 genes in invertebrates and vertebrate species arose from a common and strongly supported duplication event (Node A, Figure 5; Supplementary Figure 1). A speciation event (Node B, Figure 5) gave

rise to the CYP4V subfamily, which is the only subfamily which includes vertebrate and invertebrate species. The placement order of the CYP4V Ascidiace and Echinoidea genes was uncertain , et speciation patterns for Gastropoda, Ascidiace, Echinoidea, and Vertebrata are well established (Adoutte *et al.* 1999; Adoutte *et al.* 2000) and are placed accordingly in Figure 5. The clustering of CYP4V with invertebrate CYP4 subfamilies is similar to previous phylogenetic analyses (Fujita *et al.* 2004; Nelson 2009) .

*3.4.4 Evolution of vertebrate CYP4 subfamilies*

All vertebrate CYP4 subfamilies, except CYP4V, were placed in a second major cluster (CYP 4F/4T/4B/4A/4X/4Z, Node D, Figure 1) that was rooted by 3 gastropod and 3 *Ciona* sequences. The clustering of invertebrate sequences with these subfamilies has not been previously reported, although perhaps not surprising in light of the tentative nomenclature for the *Ciona* sequences (CYP4F). While our phylogenetic reconstructions do not necessarily support the assignment of these sequences to the CYP4F subfamily, or any other existing vertebrate CYP4 subfamily, certainly there are invertebrate CYP4 homologs to more than just CYP4V genes, as had been previously thought.

A strongly supported duplication event (Node G, Figure 6) gave rise to the CYP4F vertebrate clade and the remaining CYP4 vertebrate subfamilies. A second duplication event (Node E, Figure 6) gave rise to the CYP4F and mammalian CYP4F22 genes from the ancestral CYP4F gene, which is in agreement with a previous study (Nelson *et al.* 2004). The CYP4F22 subclade is basal to the CYP4F subfamily, and only

found in extant mammalian species suggesting a loss of CYP4F22 in all other vertebrate classes (Figure 6).

The actinopterygian CYP4T genes are basal in the more recently evolved cluster of CYP4 subfamilies (Node H, Figure 7), as was seen previously (Fujita *et al.* 2004; Nelson 2003). Speciation events gave rise to CYP4T genes in amphibian and CYP4B genes in avian species. Strong support was not found for the placement of the avian CYP4B sequences that appear ancestral to the CYP4B mammalian clade, yet there is good support for inclusion of CYP4B avian clade with the CYP4B and CYP4T clusters. The nomenclature for avian sequences was inconsistent since one sequence did not have clear nomenclature (CYP4A/4B retrieved from the P450 homepage), yet it clustered with the two CYP4B avian sequences (from GenBank) distinct from the CYP4B and CYP4A mammalian subfamily clades (Supplementary Figure 1). The CYP4B and the CYP4T genes will require further study in order to resolve the nomenclature and the currently identified topology. The CYP4A/4X/4Z cluster was strongly supported (Node I, Figure 7) and the topology resolves prior inconsistencies, and were found to match the topology from two studies that identified CYP4X and CYP4Z as sister subfamilies that arose from the same duplication with CYP4A as the ancestral subfamily (Nelson *et al.* 2004; Thomas 2007). Interestingly, CYP4Z has only been identified in humans and exhaustive searching in the mouse genome revealed no CYP4Z gene and the homologous genomic region was missing (Hsu *et al.* 2007; Nelson *et al.* 2004). The recent evolution of the CYP4 family in the mammalian species likely resulted in specific functional characterization and expression preferences between the different enzymes.

### *3.4.5 DIVERGE type I functional analysis*

The CYP4 subfamilies were analyzed for regions with high rates of evolutionary divergence, because regions with significant divergence may be associated with altered function or substrate selectivity. Regions with high rates of evolutionary divergence have been seen in analyses of CYP3A (McArthur *et al.* 2003) and CYP1 (Goldstone *et al.* 2007) proteins using this approach. In an analysis of the CYP2 family, a functionally diverse and very large CYP family, high rates of evolutionary divergence were seen to extend across the protein for most subfamily comparisons (Kirischian *et al.* 2011); yet clusters of high evolutionary rates of functional divergence were found in CYP3As and these regions did overlap with the hypothesized SRS yet they were not restricted to these regions (McArthur *et al.* 2003). Yet CYP19, a protein with nearly identical function across vertebrate species, had no functional divergence across vertebrate taxa (Wilson *et al.* 2005). Collectively, these studies suggest this approach may be useful in some CYP families to identify regions within the protein that may be functionally important or determine functional differences across subfamilies or taxonomic groups.

In general, most of the subfamily group comparisons did not show any distinct high divergence patterns throughout the alignment (Figure 2); similar to what has been seen in analyses of the CYP2 family (Kirischian *et al.* 2011). Approximately one quarter of the pairwise type I comparisons were not statistically significant (Figure 2; Supplementary Table 2), suggesting that these subfamilies may be functionally similar. For example, there was no statistically significant functional divergence identified

between CYP4V proteins in mammalians versus fish (Supplementary Table 2), suggesting that function is likely conserved across all CYP4V genes in vertebrates. Surprisingly, no statistical significance was identified for comparison of fungi & choanoflagellates with the CYP4F, CYP4V, CYP4X and CYP4B (mam) vertebrate groups (Supplementary Table 2). The functional significance of this is unclear.

Throughout the alignment, low rates of evolutionary divergence were detected across all pairwise comparisons in several helices (B, B', J', K', K'') and a region on the C-terminus (Figure 2), suggesting that these regions are composed of amino acids that are highly conserved in CYP4 genes. These helices are associated with highly or moderately variable CYP regions (Graham and Peterson 1999; Hasemann *et al.* 1995), thus these patterns are likely specific to the CYP4 family function. In a study of CYP2 functional divergence, helices B, B' and C-terminus of J, and a region at the N-terminus showed low or no divergence throughout all pairwise comparisons (Kirischian *et al.* 2011); suggesting that helices B and B' may have lower variability than previously thought in at least some CYP families. Helix B and B' are regions important for the access channel for when the substrate enters the enzyme.

Interestingly, the CYP4X subfamily had very high divergence and the CYP4F22 mammalian group had low/no divergence in multiple (10/12) comparisons across the entire protein; suggesting that there might be underlying functional similarity or that the rates of evolution were not statistically significant from each other. In the CYP4X group, high evolutionary rates of functional divergence were found throughout the CYP4

alignment (Figure 2) which may imply that the more recently evolved subfamilies have been under positive functional selection.

### *3.4.6 DIVERGE type II functional analysis*

CYP4 subfamily sister groups were analyzed for radical biochemical changes thus identifying specific residues that are conserved and may be essential for metabolic activity within one group yet different in another. The identified radical biochemical changes did not cluster only in the proposed active site regions and in most comparisons they were scattered across the protein (data not shown). Analyses of radical biochemical changes across CYP2 subfamilies identified clustered residues in the active site of the protein, suggesting amino acid changes important for substrate specificity in certain CYP2 subfamilies (Kirischian *et al.* 2011). Contrary to the CYP2 analyses (Kirischian *et al.* 2011), the sites with radical biochemical changes were few, except for comparisons amongst the CYP4A/4X/4Z subfamilies.

### *3.4.7 Vertebrate CYP4 gene expression*

The collected EST data was statistically assessed and a value of 6 TPM was determined as a cut of value (in the lower 5% of EST distribution), similar cut-off values have been used previously (Maher *et al.* 2006; Schmitt *et al.* 1999). Consideration of library size is essential, since each library should be sequenced deeply enough to provide reasonable data for low to moderately expressed genes. The library EST cut-off values were set to a minimum of 10,000 ESTs (Schmitt *et al.* 1999). Those genes with TPM values of 6 and above in such libraries (≥10,000 ESTs), likely represents expression of

frequently expressed genes; genes with less than 6 TPM are likely to have no or very low expression. In general, using ESTs to detected transcription levels of unknown genes can be quite powerful in identifying putative expression profiles thus furthering the understanding of gene expression and possibly the functional specificity (Nagaraj *et al.* 2007).

Since knowledge of CYP4 gene expression is primarily limited to mammals, utilizing the EST database from species across the 4 vertebrate classes may begin to shed light on the function of homologous genes. In mammals, the metabolic activity of CYP4 proteins varies from broad to specialized activity depending on the subfamily to which they belong (Hardwick 2008; Simpson 1997). CYP enzymes associated with chemical defense are typically expressed in organs involved in absorption, metabolism and excretion of foreign compounds including lung, intestine, liver, and kidney. CYP enzymes primarily involved in production and metabolism of endogenous signaling molecules would not be restricted to organs involved in absorption and excretion. We would hypothesize that if tissue expression patterns of homologous genes are conserved across vertebrate species, function may be conserved. Very different expression patterns across homologous genes should be indicative of functional diversity. Indeed, the *in silico* gene expression analyses showed highly variable expression in CYP4X and across species within the CYP4V subfamilies (Figure 4; Supplementary Figure 2), suggesting functional diversity.

### *3.4.8 Vertebrate CYP4F functional analyses*

It has been hypothesized that there was one ancestral CYP4F gene that gave rise to CYP4F22 in mammals. The CYP4F43 gene in zebrafish (Goldstone *et al.* 2010) and the CYP4F39 gene in mice shared synteny with the mammalian CYP4F22 genes, suggesting a possible functional similarity (Nelson *et al.* 2004). Low or no functional divergence was found for the majority of comparisons with CYP4F22 group (Figure 2; Supplementary Table 2). Type II divergence within the CYP4F subfamily (Node E, Figure 1) was lowest with comparisons including CYP4F22 (Figure 2). There were 9 and 10 sites identified as divergent between CYP4F22 and CYP4F fish and CYP4F*, respectively (Supplementary Table 3); divergent residues were located from the center of the protein to the C-terminus. Only 2 of these sites were shared and were found in substrate binding and heme stabilizing regions of the protein (Graham and Peterson 1999). Interestingly, the highly variable F and G helices and FG loop (Graham and Peterson 1999; Hasemann *et al.* 1995) had no detected biochemical divergence, implying that the ceiling of the substrate binding region is conserved across CYP4F and CYP4F22 genes and that the substrates are likely similar in structure. Also interestingly, type I functional divergence between the CYP4F fish and CYP4F* groups identified double the number of divergent sites compared to those analyses that included CYP4F22. Divergent sites were detected in the N-terminus of the protein and small clusters of radical biochemical changes were localized to helix G, the B-B' loop, helix J' and K-K'' regions, which compose the access channel and the ceiling for substrate binding (Graham and Peterson 1999; Hasemann *et al.* 1995; Li and Poulos 1996). Functional divergence is likely higher between the fish CYP4F and mammalian CYP4F* genes.

The expression of the human CYP4F22 gene was below the cutoff limits within the selected common organs (Figure 4B); yet low levels of expression were found in the skin, prostate, and spleen (9, 15, and 18 TPM, respectively; data not shown). The EST expression patterns of CYP4F22 appear more distinct from other mammalian CYP4F genes than expression patterns of CYP4 genes in non-mammalian vertebrates. Yet, some similarities in expression sites were found; human CYP4F2 and CYP4F3 gene expression was identified in skin and spleen, respectively. RT-PCR analyses identified CYP4F22 gene expression in skin, kidney, brain, testes, placenta, bone marrow (Lefevre *et al.* 2006), a more diverse expression pattern than found with EST data in this study. Mammalian CYP4F* genes were expressed in similar organs and at similar levels as CYP4F genes from fish and amphibian species (Figure 4B). Although no CYP4F EST expression was identified in the zebrafish liver library, zebrafish liver microarray data showed expression in both male and females (Wilson J.Y., unpublished data). The CYP4F genes, excluding CYP4F22, were expressed in all vertebrates in the brain, kidney, liver and possibly intestines (Figure 4B; Fujita *et al.* 2004; Hardwick 2008; Kalsotra *et al.* 2002; Kikuta *et al.* 2002; Sabourault *et al.* 1998) suggesting a broad expression for most CYP4F genes.

The functional roles of mammalian CYP4F* enzymes include the metabolism of an array of eicosanoids from leukotrienes and prostaglandin classes, long chain fatty acids, and exogenous compounds that vary in structural composition (i.e. benzphetamine ($C_{17}$), erythromycin ($C_{30}$), and ebastine ($C_{32}$); Hardwick 2008; Kalsotra and Strobel 2006; Kalsotra *et al.* 2004; Kikuta *et al.* 2002; Kikuta *et al.* 2007). CYP4F22 has been

suggested to ω-hydroxylate eicosanoids including trioxilin A3 ($C_{20}H_{34}O_5$) from the 12(R)-lipoxygenase pathway, an essential pathway for maintaining the skin permeability (Lefevre *et al.* 2006). The homologous genes CYP4F2 and CYP4F3 participate in a parallel pathway, 5(S)-lipoxygenase, where they ω-hydroxylate leukotriene B4 (Lefevre *et al.* 2006). The functional capabilities of CYP4F22 enzymes are still unknown with regards to which fatty acids or eicosanoids these genes preferentially metabolize (Hardwick 2008). The specific metabolic pathways and the physiological role of CYP4F22 is largely unknown; apart from a likely role in skin and certain skin diseases (Lefevre *et al.* 2006). In fish, little is known of the CYP4F gene function, one study suggested hydroxylation of the eicosanoids leukotriene $B_4$ as a possible function; yet the biological importance of leukotriene $B_4$ is unknown in fish (Sabourault *et al.* 1998). Thus based on the supporting lines of evolutionary and *in silico* functional divergence analyses, it is likely that CYP4F22 hydroxylates eicosanoids and has a conserved functional importance, which is likely shared with the CYP4F genes in fish. The functional divergence between CYP4F22 and CYP4F* genes is likely more subtle, distinct eicosanoids substrates appear to be metabolized by CYP4F22 and other CYP4F genes in humans, yet the patterns of expression are quite distinct suggesting divergent roles in physiological processes.

### *3.4.9 Mammalian CYP4A and CYP4X functional analyses*

The highest rates of type I functional divergence was associated with CYP4X subfamily pairwise comparisons (Figure 2). High evolutionary rates of functional

divergence were also seen with the mammalian CYP4A genes (Supplementary Table 2), yet the distribution of sites was not as dense throughout the protein as seen with CYP4X comparisons. The type II biochemical functional analyses detected largest number of sites and highest degree of radical biochemical changes within recently the evolved mammalian subfamilies, CYP4A, CYP4X, and CYP4B (Figure 3A-D). Of these three comparisons, the highest divergence in both magnitude and sites numbers was associated with the CYP4X subfamily (data not shown). These are a strong indicator that the CYP4X, CYP4A and CYP4B mammalian subfamilies likely have distinct functions. Indeed, clustering of sites with radical biochemical changes were seen between all three comparisons, yet the numbers of residues involved was larger in the CYP4X/CYP4A comparison, and the majority of sites were found in the active site regions (Figure 3). The CYP4As have been well established to favor metabolism of ω-side medium chain fatty acids (Bellamine *et al.* 2003; Okita and Okita 2001). The CYP4A11 gene was able to oxidize arachidonic acid (long chain fatty acid) to 20-HETEs (Bellamine *et al.* 2003; Savas *et al.* 2005), yet CYP4A enzymes have 10-100 times greater metabolic rates towards the linear, saturated medium fatty acids, lauric ($C_{12}$) and palmitic ($C_{16}$) acid (Hoch *et al.* 2000; Kawashima *et al.* 2000; Palmer *et al.* 1993). The metabolic capabilities of CYP4X are largely unknown, yet a recent study has identified selective oxidation by CYP4X1 of endocannabinoid anandamides, long chain fatty acids amides ($C_{22}$), that are important for the nervous and immune systems (Stark *et al.* 2008). Interestingly, human CYP4F2 was capable of anandamide metabolism (Snider *et al.* 2007) and the CYP4F subfamily is known for metabolizing long chain fatty acids,

suggesting a functional overlap between CYP4F and CYP4X that may account for the low type I divergence seen between CYP4X and CYP4F22 clusters in our study.

Expression data for CYP4A and CYP4X mammalian genes were quite distinct. CYP4A expression, supported by this and prior studies, was restricted to kidney and liver (Figure 4A; Cowart *et al.* 2002; Palmer *et al.* 1993; Savas *et al.* 2005; Simpson 1997). CYP4X expression was ubiquitous in the EST libraries and had been previously reported in liver, aorta, brain, heart, breast, colon, skin (Savas *et al.* 2005; Stark *et al.* 2008) and kidney (Al-Anizy *et al.* 2006). In our analysis and others (Savas *et al.* 2005), the highest CYP4X expression was found in the trachea (Figure 4A) although the biological importance of this had not been determined. The expression of CYP4X in tissues associated with neuronal function, such as brain, was high (Bylund *et al.* 2002; Savas *et al.* 2005; Stark *et al.* 2008) and our data agrees. CYP expression in the brain is limited; brain CYP expression was only 0.5%-2% of hepatic CYP expression (Hedlund *et al.* 2001). High levels of ω-side hydroxylation of fatty acids have been detected in the brain (Hedlund *et al.* 2001), a reaction that is presumably CYP mediated as it is detected elsewhere (Hardwick 2008; Hsu *et al.* 2007). This suggests that CYP4X1 may have a primary role in ω-side hydroxylation of fatty acids in the brain (Hedlund *et al.* 2001), likely long chain fatty acid amides such as anandamide (Stark *et al.* 2008) and a physiological role in neuronal function.

### 3.5 Conclusion

This study includes a phylogenetic analysis of the CYP4 family illustrating the evolutionary history of subfamilies from invertebrate (insects) and vertebrates. Of all CYP4 subfamilies, only CYP4V was found in invertebrate and vertebrate species and the subfamily clustered in close proximity with other invertebrate CYP4 subfamilies. The 6 vertebrate specific subfamilies clustered in a separate clade from CYP4V; only CYP4F was found in all 4 vertebrate classes. Complex topology was found for the CYP4T and CYP4B subfamilies; the CYP4T and CYP4B subfamilies shared a common ancestor but did not cluster in distinct monophyletic subfamily clades. Further consideration should be given to the nomenclature of the CYP4T and CYP4B subfamily genes, particularly those found in avian species. Strong support was seen for the placement of CYP4A as a basal subfamily to the recently evolved sister subfamilies, CYP4X and CYP4Z. Low evolutionary rates of functional divergence within the B, B',J ', K', K'' helices and a region at the end of C-terminus (Figure 2) were found in all pairwise type I functional divergence comparisons suggesting these are conserved regions in CYP4 genes. Based on functional divergence, gene expression data combined with some prior functional studies, CYP4X has been hypothesized to ω-hydroxylates long chain fatty acid amides (i.e. anandamide) and may play an important physiological role in the neuronal tissues. Both type I and II functional divergence analyses suggest that CYP4F22 genes have a similar function to other CYP4F genes, although gene expression sites were different. Functional testing of CYP4F22 clade should likely include ω-hydroxylation of long chain

fatty acids in particular eicosanoids (i.e. trioxilin A3), which are a part of an essential pathway that preserves skin permeability.

**Acknowledgements**

We would like to thank Drs. Brian Golding and Jonathon Stone for computer cluster access for Bayesian analysis and assessment of posterior probabilities of nodes via PAUP* software, respectively. This project was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant #328204 to JYW). The Department of Biology, McMaster University, provided partial support for N.K.

### 3.6 References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-5

Adoutte A, Balavoine G, Lartillot N, de Rosa R (1999) Animal evolution. The end of the intermediate taxa? Trends Genet 15:104-8

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R (2000) The new animal phylogeny: reliability and implications. Proc Natl Acad Sci U S A 97:4453-6

Al-Anizy M, Horley NJ, Kuo CW, Gillett LC, Laughton CA, Kendall D, Barrett DA, Parker T, Bell DR (2006) Cytochrome P450 Cyp4x1 is a major P450 protein in mouse brain. Febs J 273:936-47

Baer BR, Rettie AE (2006) CYP4B1: an enigmatic P450 at the interface between xenobiotic and endobiotic metabolism. Drug Metab Rev 38:451-76

Baldwin WS, Marko PB, Nelson DR (2009) The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. BMC Genomics 10:169

Bellamine A, Wang Y, Waterman MR, Gainer JV, 3rd, Dawson EP, Brown NJ, Capdevila JH (2003) Characterization of the CYP4A11 gene, a second CYP4A gene in humans. Arch Biochem Biophys 409:221-7

Bylund J, Hidestrand M, Ingelman-Sundberg M, Oliw EH (2000) Identification of CYP4F8 in human seminal vesicles as a prominent 19-hydroxylase of prostaglandin endoperoxides. J Biol Chem 275:21844-9

Bylund J, Zhang C, Harder DR (2002) Identification of a novel cytochrome P450, CYP4X1, with unique localization specific to the brain. Biochem Biophys Res Commun 296:677-84

Choudhary D, Jansson I, Stoilov I, Sarfarazi M, Schenkman JB (2005) Expression patterns of mouse and human CYP orthologs (families 1-4) during development and in different adult tissues. Arch Biochem Biophys 436:50-61

Cowart LA, Wei S, Hsu MH, Johnson EF, Krishna MU, Falck JR, Capdevila JH (2002) The CYP4A isoforms hydroxylate epoxyeicosatrienoic acids to form high affinity peroxisome proliferator-activated receptor ligands. J Biol Chem 277:35105-12

Cui X, Kawashima H, Barclay TB, Peters JM, Gonzalez FJ, Morgan ET, Strobel HW (2001) Molecular cloning and regulation of expression of two novel mouse CYP4F genes: expression in peroxisome proliferator-activated receptor alpha-

deficient mice upon lipopolysaccharide and clofibrate challenges. J Pharmacol Exp Ther 296:542-50

Cui X, Nelson DR, Strobel HW (2000) A novel human cytochrome P450 4F isoform (CYP4F11): cDNA cloning, expression, and genomic structural characterization. Genomics 68:161-6

Davies L, Williams DR, Aguiar-Santana IA, Pedersen J, Turner PC, Rees HH (2006) Expression and down-regulation of cytochrome P450 genes of the CYP4 family by ecdysteroid agonists in Spodoptera littoralis and Drosophila melanogaster. Insect Biochem Mol Biol 36:801-7

Dunkov BC, Rodriguez-Arnaiz R, Pittendrigh B, ffrench-Constant RH, Feyereisen R (1996) Cytochrome P450 gene clusters in Drosophila melanogaster. Mol Gen Genet 251:290-7

Feyereisen R (1999) Insect P450 enzymes. Annu Rev Entomol 44:507-33

Feyereisen R (2006) Evolution of insect P450. Biochem Soc Trans 34:1252-5

Fujita Y, Ohi H, Murayama N, Saguchi K, Higuchi S (2004) Identification of multiple cytochrome P450 genes belonging to the CYP4 family in Xenopus laevis: cDNA cloning of CYP4F42 and CYP4V4. Comp Biochem Physiol B Biochem Mol Biol 138:129-36

Goldstone JV, Goldstone HM, Morrison AM, Tarrant A, Kern SE, Woodin BR, Stegeman JJ (2007) Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family. Mol Biol Evol 24:2619-31

Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jonsson ME, Nelson DR, Stegeman JJ (2010) Identification and developmental expression of the full complement of Cytochrome P450 genes in Zebrafish. BMC Genomics 11:643

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

Graham SE, Peterson JA (1999) How similar are P450s and what can their differences teach us? Arch Biochem Biophys 369:24-9

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500-1

Hardwick JP (2008) Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases. Biochem Pharmacol 75:2263-75

Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure 3:41-62

Hedlund E, Gustafsson JA, Warner M (2001) Cytochrome P450 in the brain; a review. Curr Drug Metab 2:245-63

Hoch U, Zhang Z, Kroetz DL, Ortiz de Montellano PR (2000) Structural determination of the substrate specificities and regioselectivities of the rat and human fatty acid omega-hydroxylases. Arch Biochem Biophys 373:63-71

Hsu MH, Savas U, Griffin KJ, Johnson EF (2007) Human cytochrome p450 family 4 enzymes: function, genetic variation and regulation. Drug Metab Rev 39:515-38

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4

Jin R, Koop DR, Raucy JL, Lasker JM (1998) Role of human CYP4F2 in hepatic catabolism of the proinflammatory agent leukotriene B4. Arch Biochem Biophys 359:89-98

Kalsotra A, Anakk S, Boehme CL, Strobel HW (2002) Sexual dimorphism and tissue specificity in the expression of CYP4F forms in Sprague Dawley rats. Drug Metab Dispos 30:1022-8

Kalsotra A, Strobel HW (2006) Cytochrome P450 4F subfamily: at the crossroads of eicosanoid and drug metabolism. Pharmacol Ther 112:589-611

Kalsotra A, Turman CM, Kikuta Y, Strobel HW (2004) Expression and characterization of human cytochrome P450 4F11: Putative role in the metabolism of therapeutic drugs and eicosanoids. Toxicol Appl Pharmacol 199:295-304

Kawashima H, Naganuma T, Kusunose E, Kono T, Yasumoto R, Sugimura K, Kishimoto T (2000) Human fatty acid omega-hydroxylase, CYP4A11: determination of complete genomic sequence and characterization of purified recombinant protein. Arch Biochem Biophys 378:333-9

Kikuta Y, Kusunose E, Kusunose M (2002) Prostaglandin and leukotriene omega-hydroxylases. Prostaglandins Other Lipid Mediat 68-69:345-62

Kikuta Y, Mizomoto J, Strobel HW, Ohkawa H (2007) Expression and physiological function of CYP4F subfamily in human eosinophils. Biochim Biophys Acta 1771:1439-45

Kirischian N, McArthur AG, Jesuthasan C, Krattenmacher B, Wilson JY (2011) Phylogenetic and Functional Analysis of the Vertebrate Cytochrome P450 2 Family. J Mol Evol 72:56-71

Lefevre C, Bouadjar B, Ferrand V, Tadini G, Megarbane A, Lathrop M, Prud'homme JF, Fischer J (2006) Mutations in a new cytochrome P450 gene in lamellar ichthyosis type 3. Hum Mol Genet 15:767-76

Lewis DF (2002) Homology modelling of human CYP2 family enzymes based on the CYP2C5 crystal structure. Xenobiotica 32:305-23

Li A, Jiao X, Munier FL, Schorderet DF, Yao W, Iwata F, Hayakawa M, Kanai A, Shy Chen M, Alan Lewis R, Heckenlively J, Weleber RG, Traboulsi EI, Zhang Q, Xiao X, Kaiser-Kupfer M, Sergeev YV, Hejtmancik JF (2004) Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. Am J Hum Genet 74:817-26

Li C, Orti G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol 7:44

Li H, Poulos TL (1996) Conformational dynamics in cytochrome P450-substrate interactions. Biochimie 78:695-9

Liu Y, Wang J, Liu Y, Zhang H, Xu M, Dai J (2009) Expression of a novel cytochrome P450 4T gene in rare minnow (Gobiocypris rarus) following perfluorooctanoic acid exposure. Comp Biochem Physiol C Toxicol Pharmacol 150:57-64

Maddison DR, Maddison WP (2000) MacClade version 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland Massachusetts

Maher C, Stein L, Ware D (2006) Evolution of Arabidopsis microRNA families through duplication events. Genome Res 16:510-9

McArthur AG, Hegelund T, Cox RL, Stegeman JJ, Liljenberg M, Olsson U, Sundberg P, Celander MC (2003) Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. J Mol Evol 57:200-11

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8:6-21

Nakano M, Kelly EJ, Rettie AE (2009) Expression and characterization of CYP4V2 as a fatty acid omega-hydroxylase. Drug Metab Dispos 37:2119-22

Nelson DR (1998) Metazoan cytochrome P450 evolution. Comp Biochem Physiol C Pharmacol Toxicol Endocrinol 121:15-22

Nelson DR (1999) Cytochrome P450 and the individuality of species. Arch Biochem Biophys 369:1-10

Nelson DR (2003) Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys 409:18-24

Nelson DR (2009) The Cytochrome P450 Homepage. Human Genomics 4:59-65

Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1-18

Okita RT, Okita JR (2001) Cytochrome P450 4A fatty acid omega hydroxylases. Curr Drug Metab 2:265-81

Palmer CN, Richardson TH, Griffin KJ, Hsu MH, Muerhoff AS, Clark JE, Johnson EF (1993) Characterization of a cDNA encoding a human kidney, cytochrome P-450 4A fatty acid omega-hydroxylase and the cognate enzyme expressed in Escherichia coli. Biochim Biophys Acta 1172:161-6

Prasad AB, Allard MW, Green ED (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25:1795-808

Ren F, Tanaka H, Yang Z (2005) An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol 54:808-18

Rewitz KF, Styrishave B, Lobner-Olsen A, Andersen O (2006) Marine invertebrate cytochrome P450: emerging insights from vertebrate and insects analogies. Comp Biochem Physiol C Toxicol Pharmacol 143:363-81

Rieger MA, Ebner R, Bell DR, Kiessling A, Rohayem J, Schmitz M, Temme A, Rieber EP, Weigle B (2004) Identification of a novel mammary-restricted cytochrome P450, CYP4Z1, with overexpression in breast carcinoma. Cancer Res 64:2357-64

Rogers J, Wall R (1980) A mechanism for RNA splicing. Proc Natl Acad Sci U S A 77:1877-9

Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, Oxbrow AK, Lewis CJ, Tennant MG, Modi S, Eggleston DS, Chenery RJ, Bridges AM (2006) Crystal structure of human cytochrome P450 2D6. J Biol Chem 281:7614-22

Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF (2008) A phylogenomic investigation into the origin of metazoa. Mol Biol Evol 25:664-72

Sabourault C, Berge J, Lafaurie M, Girard JP, Amichot M (1998) Molecular cloning of a phthalate-inducible CYP4 gene (CYP4T2) in kidney from the sea bass, Dicentrarchus labrax. Biochem Biophys Res Commun 251:213-9

Savas U, Hsu MH, Griffin KJ, Bell DR, Johnson EF (2005) Conditional regulation of the human CYP4X1 and CYP4Z1 genes. Arch Biochem Biophys 436:377-85

Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B, Rosenthal A (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res 27:4251-60

Simpson AE (1997) The cytochrome P450 4 (CYP4) family. Gen Pharmacol 28:351-9

Snider NT, Kornilov AM, Kent UM, Hollenberg PF (2007) Anandamide metabolism by human liver and kidney microsomal cytochrome p450 enzymes to form hydroxyeicosatetraenoic and epoxyeicosatrienoic acid ethanolamides. J Pharmacol Exp Ther 321:590-7

Srivastava H, Sharma M, Dixit J, Das A (2010) Evolutionary insights into insecticide resistance gene families of Anopheles gambiae. Infect Genet Evol 10:620-8

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758-71

Stark K, Dostalek M, Guengerich FP (2008) Expression and purification of orphan cytochrome P450 4X1 and oxidation of anandamide. Febs J 275:3706-17

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3:e67

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876-82

Tijet N, Helvig C, Feyereisen R (2001) The cytochrome P450 gene superfamily in Drosophila melanogaster: annotation, intron-exon organization and phylogeny. Gene 262:189-98

Waxman DJ (1999) P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptors CAR, PXR, and PPAR. Arch Biochem Biophys 369:11-23

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9

Whittingham LA, Slikas B, Winkler DW, Sheldon FH (2002) Phylogeny of the tree swallow genus, Tachycineta (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Mol Phylogenet Evol 22:430-41

Wilson JY, McArthur AG, Stegeman JJ (2005) Characterization of a cetacean aromatase (CYP19) and the phylogeny and functional conservation of vertebrate aromatase. Gen Comp Endocrinol 140:74-83

**Table 3.1. CYP4 sequences by source and species.** Sequences were collected during *de novo* annotation of complete genomes (*de novo*) or retrieved from the P450 homepage, Ensembl and NCBI (GenBank). Species with *de novo* sequences or from the P450 homepage represent the species with complete CYP4 complements. GenBank sequences were retrieved with exhaustive BLAST searching and do not represent all subfamilies for the species. See methods for detail.

| Class | Species | NCBI/Ensembl | P450 Homepage | *De novo* annotation addition |
|---|---|---|---|---|
| *Mammalia* | **Cow** | | | 9 |
| | **Dog** | | | 8 |
| | **Giant Panda** | 2 | | |
| | **Goat** | 1 | | |
| | **Horse** | 8 | | |
| | **Human** | | 12 | |
| | **Koala** | 1 | | |
| | **Minke whale** | 2 | | |
| | **Opossum** | 6 | | |
| | **Pig** | 2 | | |
| | **Platypus** | 2 | | |
| | **Rabbit** | 12 | | |
| | **Sheep** | 1 | | |
| *Aves* | **Chicken** [1,2] | 1 | 2 | |
| | **Zebra finch** | 2 | | |
| *Amphibia* | **Frog (*X.lavies*)** [1] | 6 | | |
| | **Frog (*X.tropicalis*)** | 5 | | |
| *Actinopterygii* | **European seabass** | 1 | | |
| | **Fugu** | | 3 | |
| | **Green pufferfish** | 1 | | |
| | **Medaka** | | | 3 |
| | **Salmon** | 1 | | |
| | **Stickleback** | | | 3 |
| | **Zebrafish** [3] | 2 | 2 | |
| *Ascidiacea* | **Ciona** | 4 | | |
| *Echinoidea* | **Sea Urchin** | 4 | | |
| *Gastropoda* | **Snail** | 4 | | |
| *Insecta* | **Drosophila** | | 22 | |

[1]- Frog sequences were checked for high sequence identity to human and zebrafish genes in the same subfamilies. These sequences were composed of ~500 amino acids, start and stop codons.

[2]-Chicken sequences were retrieved from GenBank and blasted against those on the P450 Homepage; 2 sequences matched sequences on the P450 server and those sequences were used for the analyses.

[3]- Zebrafish CYP4T and CYP4F sequences are from the P450 homepage, and 2 CYP4V sequences were found in Ensembl but were not included on the P450 homepage.

## 3.8 Figures

**Figure 3.1. Cytochrome P450 family 4 phylogeny.** A cartoon representation of the evolutionary relationships between the CYP4 subfamilies from invertebrates and vertebrate species. The phylogeny was determined by maximum likelihood and a WAG+I+G (Ren *et al.* 2005; Whelan and Goldman 2001) model (see materials and methods for details of phylogenetic methods). The consensus tree from a Bayesian phylogenetic analysis was in agreement for most clusters with the tree shown. The phylogenetic support (bootstrapping values / posterior probabilities) for the labeled nodes are: A (98/1.00), B (82/0.97), C (92/1.00), D (100/1.00), E (100/1.00), F (86/1.00), G (81/1.00), H (99/1.00), I (69/1.00), and J (99/1.00). A detailed phylogenetic reconstruction with phylogenetic support for all nodes can be found in Supplementary Figure 1. The nodes that include CYP4F22 sequences are marked with a single asterix (*). Those nodes that include CYP4 putative genes that lack subfamily assignments are indicated with a CYP4**. The *Drosophila* CYP4 genes, which represent 10 insect CYP4 subfamilies, are marked with CYP4***.

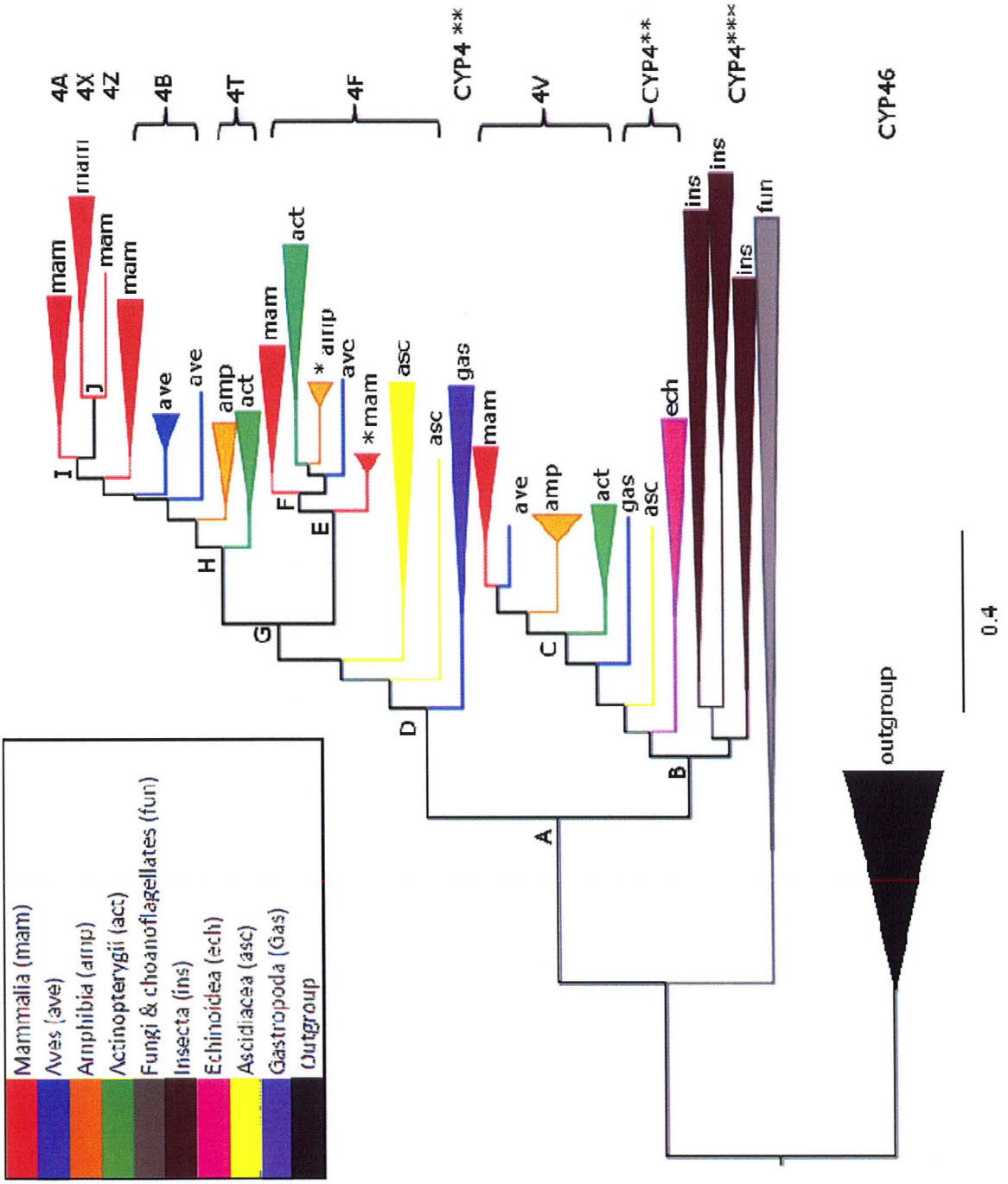

Mammalia (mam)
Aves (ave)
Amphibia (amp)
Actinopterygii (act)
Fungi & choanoflagellates (fun)
Insecta (ins)
Echinoidea (ech)
Ascidiacea (asc)
Gastropoda (Gas)
Outgroup
4A
4X
4Z
4B
4T
4F
CYP4 **
4V
CYP4**
CYP4***
CYP46
A
B
C
D
E
F
G
H
I
J
outgroup
0.4

**Figure 3.2. A hierarchical cluster analysis (heat map) of type I functional divergence of CYP4 pairwise comparisons.** A graphical representation of a model CYP4 gene, with appropriately scaled helices (A-L), is displayed above the heat map. The heat map illustrates site-specific evolutionary rates of functional divergence in 76 CYP4 pairwise comparisons. Regions with gaps or low homology in the alignment and not utilized for DIVERGE or phylogenetic analyses are shaded in grey. Black regions indicate amino acid positions with a coefficient of evolutionary rate of functional divergence, $\theta < 0.5$; red regions represent $\theta \geq 0.5$. The heat map was scaled to the human CYP4B1 gene and reference amino acid positions for unmasked regions are labeled below. The comparisons with CYP4X and mammalian CYP4F22 genes were clustered and are labeled on the right column. CYP4F22 (mam) represents the mammalian CYP4F22 cluster.

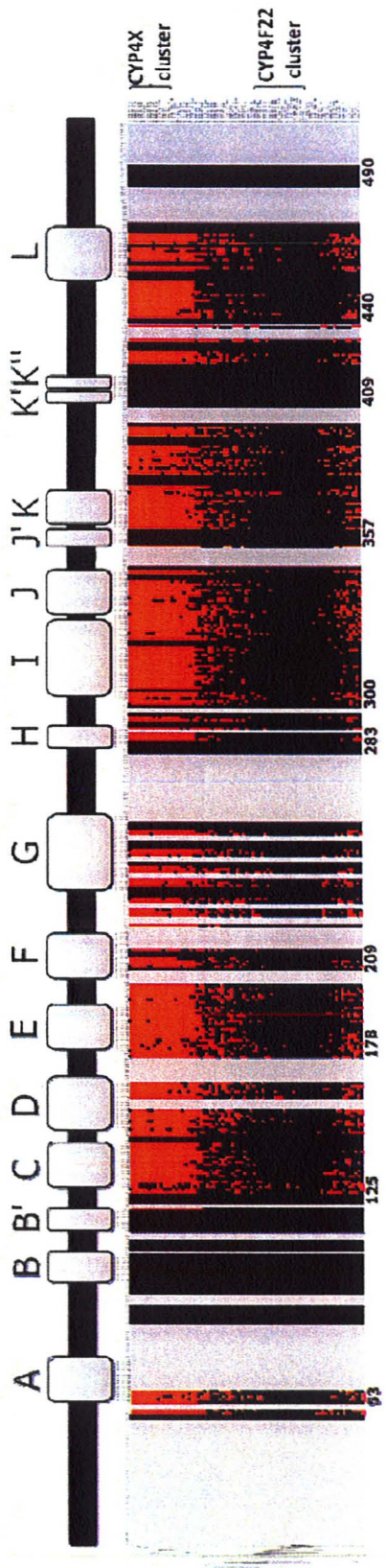

CYP4X cluster
CYP4F22 cluster
A
B
B'
C
D
E
F
G
H
I
J
J'K
K'K''
L
490
440
409
357
300
283
209
178
125
93

**Figure 3.3. Type II functional divergence in the active site between CYP4X and CYP4A genes.** DIVERGE type II functional analysis of radical biochemical changes identified 79 sites with radical changes ($\theta$>1) between genes in the CYP 4X and CYP4A mammalian subfamilies; the 32 sites with radical biochemical changes in the protein active site are shown. The sites are mapped on, with residue numbers that correspond to, the human CYP4X1 hypothetical structure (A) Helices G, F, I, J' and the FG and GH loop are labeled for orientation. (B-D) The protein structure is rotated and enlarged to clearly show changes in the N-terminus of helix F (residue 211-215), helix G (residue 243-267) and in the FG-loop (residue 228-233, panel B); helix I (residue 308-331, panel C); and between helices G-H (residue 270-284) and J' helix (residue 360-363, panel D).

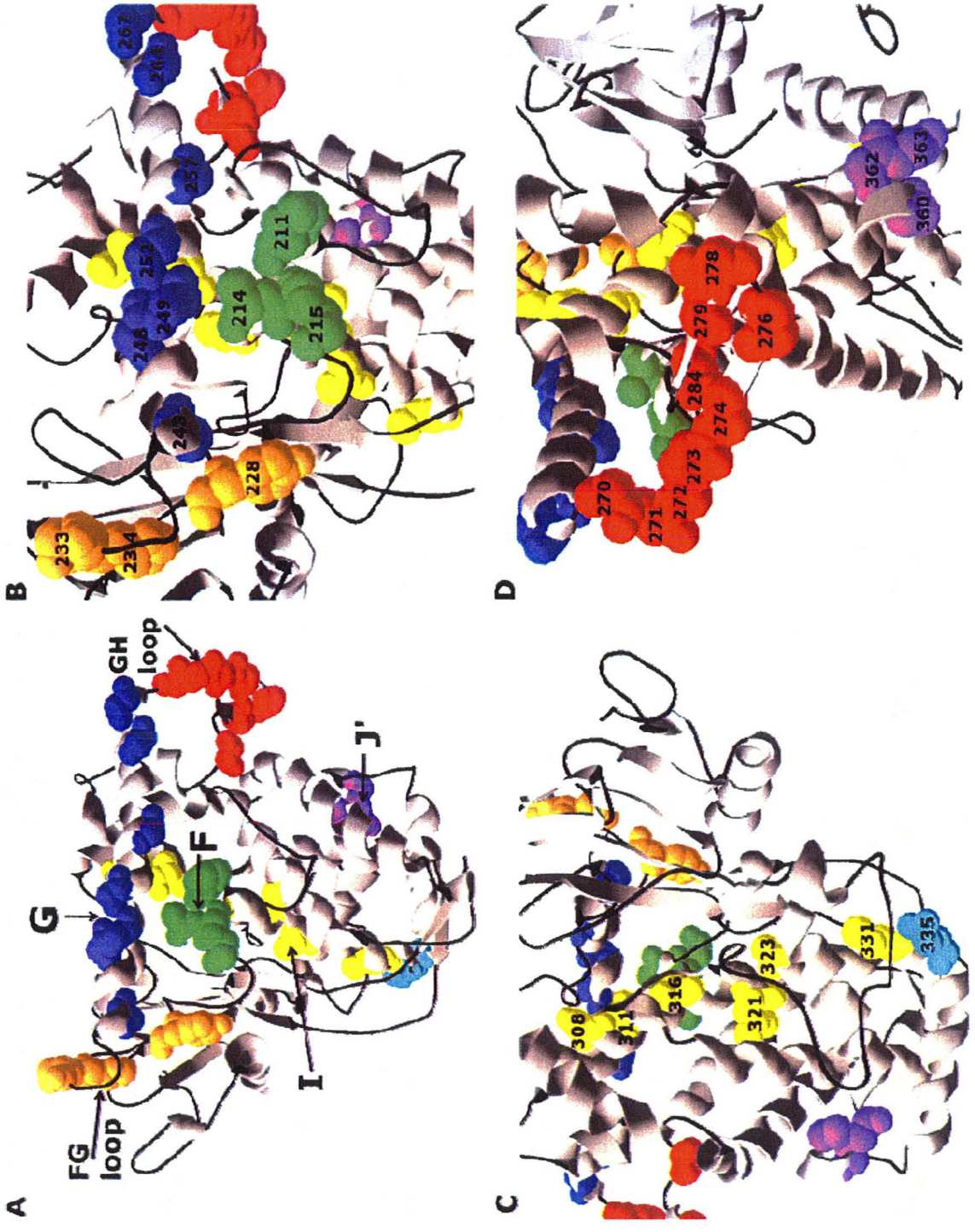
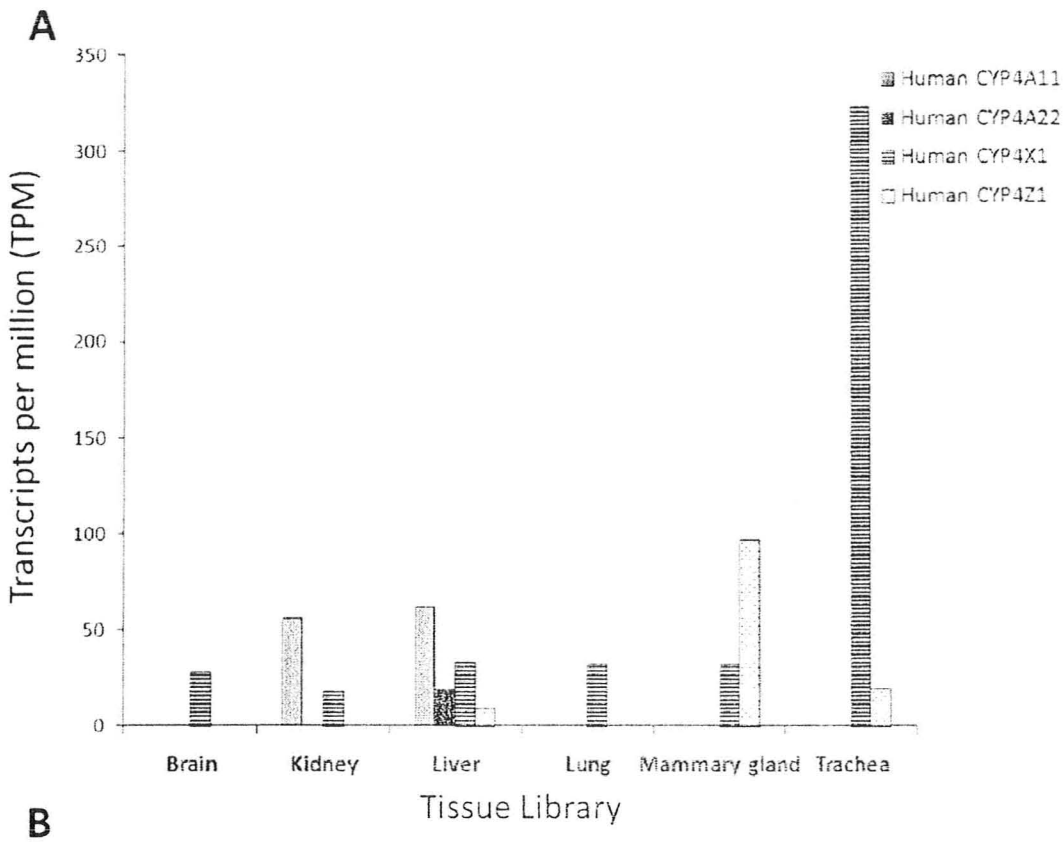

A
FG loop
G
F
GH loop
I
J'
B
233
234
228
249
214
211
215
248
249
253
257
261
267
C
308
311
316
321
323
331
335
D
270
271
272
273
274
284
276
279
278
360
362
363

**Figure 3.4. Digital gene expression of genes in the vertebrate CYP 4A, 4X, 4Z, and 4F subfamilies.** Gene expression levels of CYP4 genes are provided, in transcripts per million (TPM), from tissue libraries containing >10,000 ESTs for vertebrate species. Libraries that were not available or had less than 10,000 ESTs for the species of interest are denoted with an X. The digital gene expression patterns are shown in (A) for human genes in CYP4A, CYP4X, and CYP4Z subfamilies and (B) for human, frog and zebrafish genes in the CYP4F subfamily. No EST data was available for CYP4F in chicken.


A
Transcripts per million (TPM)
350
300
250
200
150
100
50
0
Human CYP4A11
Human CYP4A22
Human CYP4X1
Human CYP4Z1
Brain
Kidney
Liver
Lung
Mammary gland
Trachea
Tissue Library

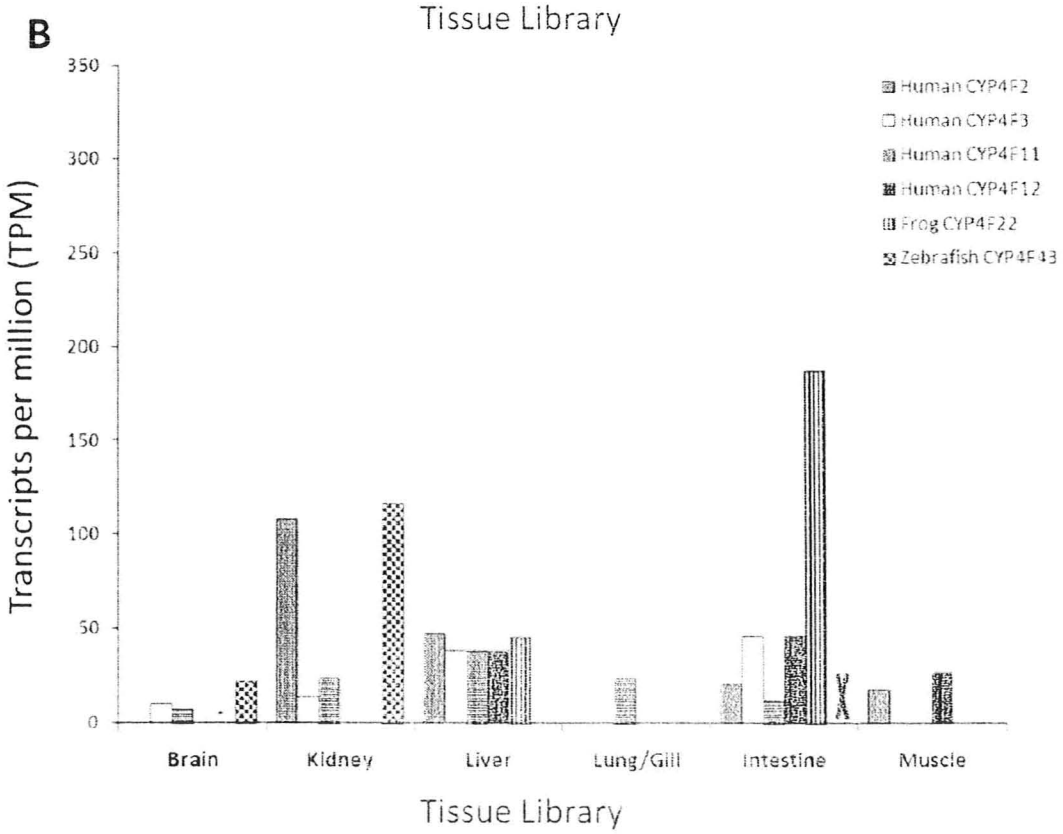
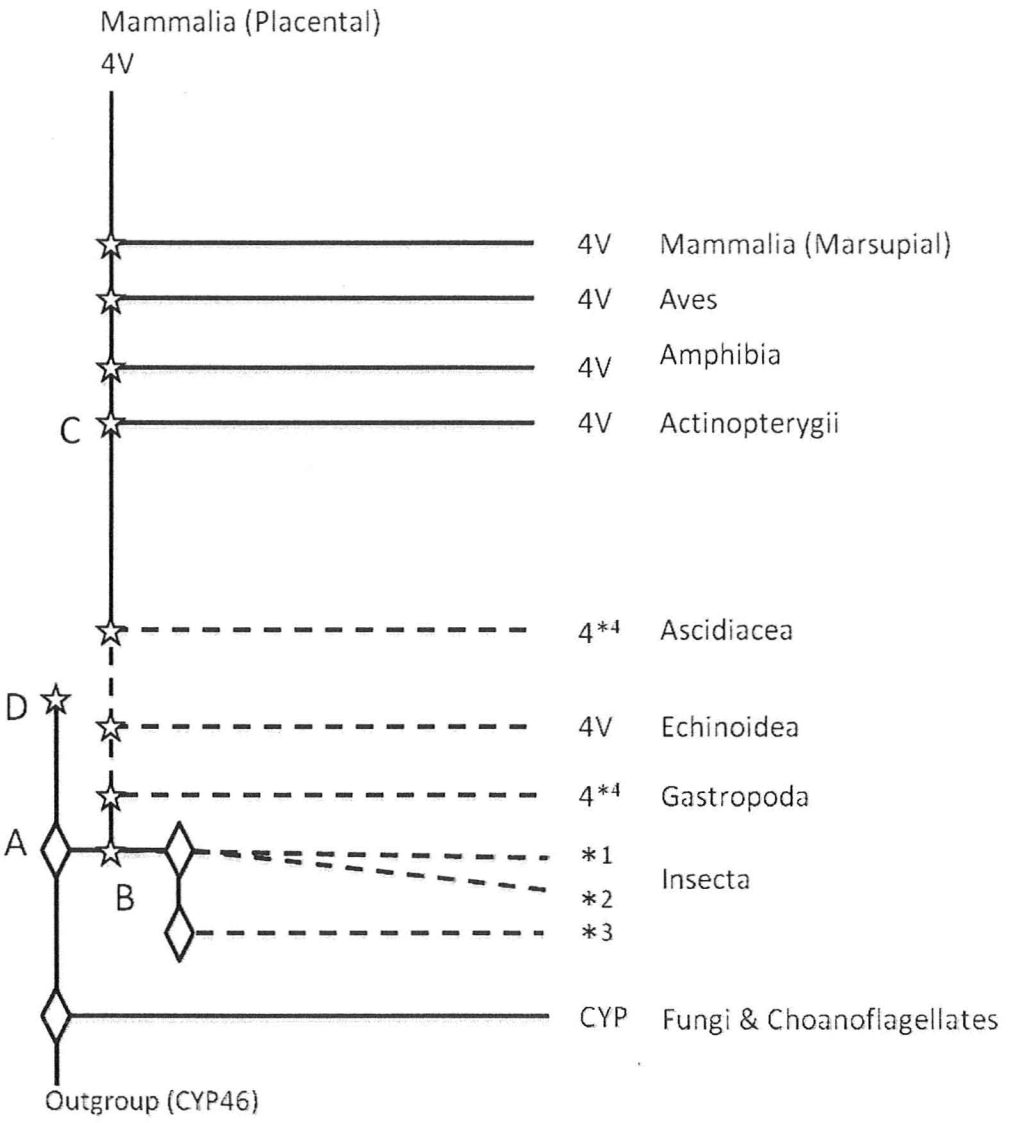

B
Transcripts per million (TPM)
350
300
250
200
150
100
50
0
Human CYP4F2
Human CYP4F3
Human CYP4F11
Human CYP4F12
Frog CYP4F22
Zebrafish CYP4F43
Brain
Kidney
Liver
Lung/Gill
Intestine
Muscle
Tissue Library

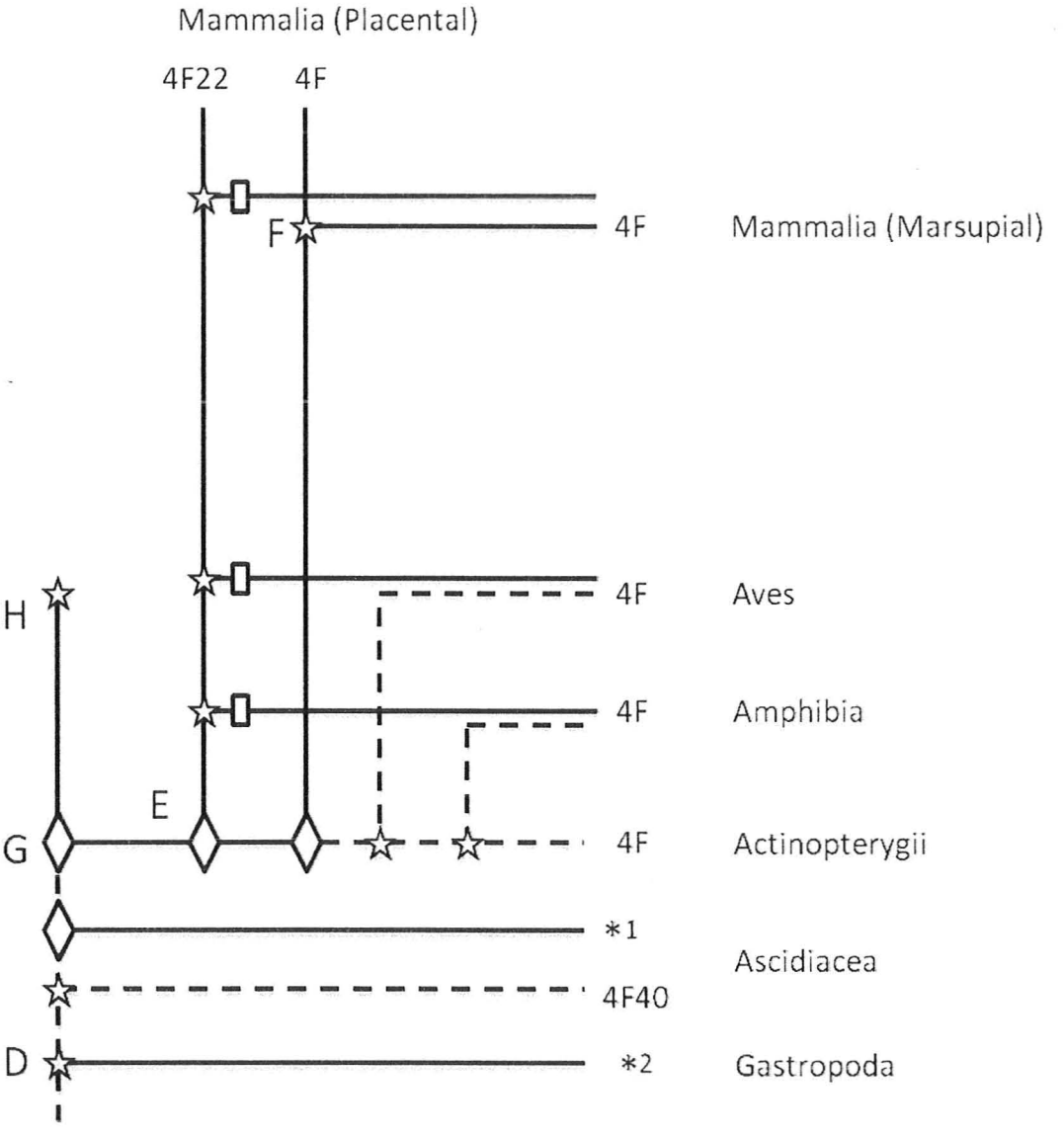
**Figure 3.5. Speciation patterns and gene duplications in the CYP4 invertebrate and CYP4V subfamilies.** The evolution of vertebrate and invertebrate CYP4 subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate and invertebrate species is symbolized by speciation patterns ( ★ ). Dashed lines represent areas of topology with weak support (see Supplementary Figure 1). Node are labeled according to those in Figure 1.




Fungi & Choanoflagellates

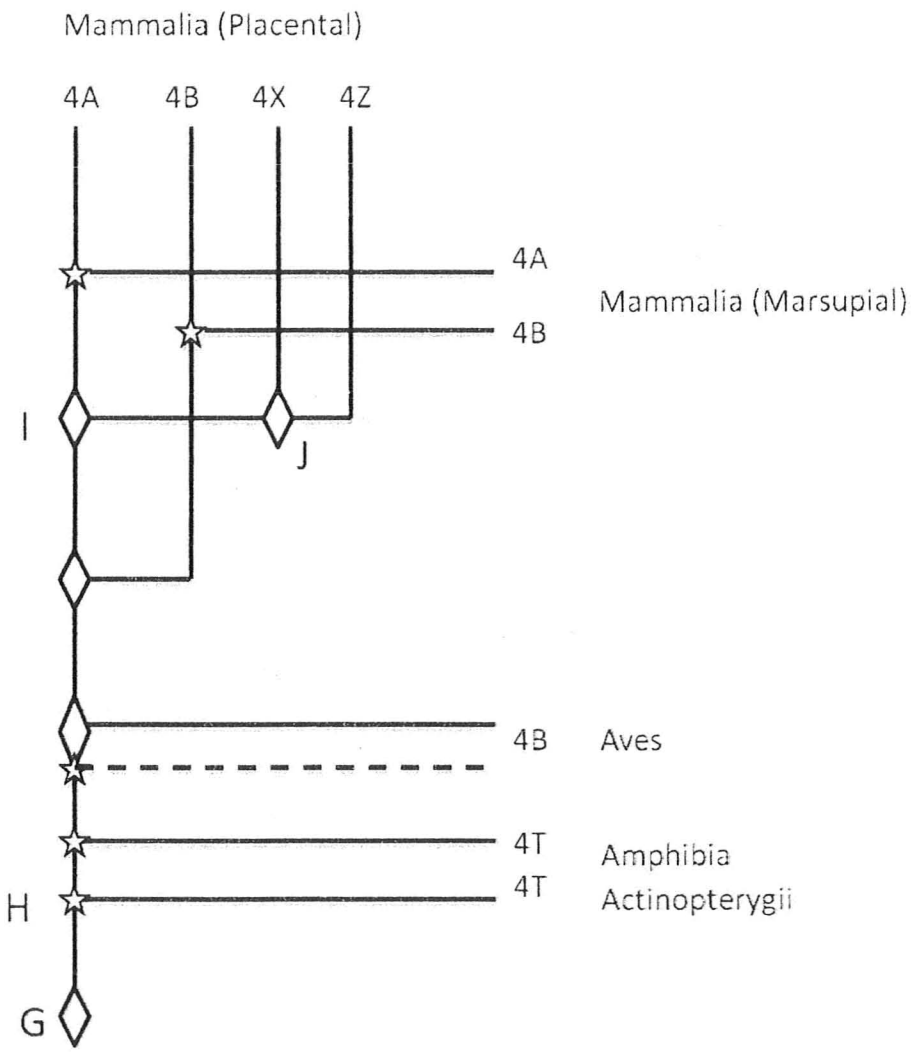
*1 Drosophila CYP 4D, 4AE
*2 Drosophila CYP 4E, 4S, 4AD
*3 Drosophila CYP 4C, 4AA, 4G15, 4G1, 4P, 4AC
*4 CYP4 putative

**Figure 3.6. Speciation patterns and gene duplications in the CYP4F subfamily.** The evolution of vertebrate and invertebrate CYP4 subfamilies are symbolized by duplication ( ) and gene loss ( ) events, strongly associated with the CYP4F22 genes in vertebrate species (excluding mammals). Diversification of vertebrate and invertebrate species is symbolized by speciation patterns ( ★ ). Dashed lines represent areas of topology with weak support (see Supplementary Figure 1). Node are labeled according to those in Figure 1.




*1 Ciona CYP 4F17 and 4F22 sequences
*2 Snail CYP4 putative sequences

**Figure 3.7. Speciation patterns and gene duplications in the vertebrate CYP 4T, 4B, 4A, 4X, and 4Z subfamilies.** The evolution of vertebrate CYP4 subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ★ ). Dashed lines represent areas of topology with weak support. Nodes are labeled according to those in Figure 1.


Mammalia (Placental)
4A
4B
4X
4Z
4A
Mammalia (Marsupial)
4B
I
J
4B
Aves
4T
Amphibia
4T
Actinopterygii
H
G

# CHAPTER 4:

# DISCUSSION

## 4.1 Phylogenetic Analyses

The CYP superfamily metabolizes endogenous and exogenous compounds and has diversified into all domains of life (Nelson 1998). The first four families of the CYP superfamily play an important role of metabolizing a large array of exogenous and endogenous compounds. In vertebrates, the evolutionary relationship outside of mammalian species in the expanded families 2 and 4 is unclear, yet their functional roles are suggested to be essential towards endogenous and exogenous compound metabolism (CYP2) and ω-hydroxylation of fatty acids (CYP4).

This thesis examines the evolutionary relationships and *in silico* functional divergence in the CYP2 and CYP4 families. For both studies, the phylogenies were reconstructed using robust methods, maximum likelihood and Bayesian inference (Huelsenbeck *et al.* 2001). The focus was to develop a better understanding of the evolutionary relationships between the vertebrate CYP subfamilies, in particular between the mammalian and actinopterygian species, and to raise functional hypotheses for the CYP genes where functional data was lacking. CYP subfamilies typically have a one-to-one ortholog subfamily relationship across vertebrates (Nelson 2003), yet CYP2 subfamilies showed species specific subfamily expansions (Figure 1, Chapter 2) as did the CYP4 families, to a lesser extent (Chapter 3). The two phylogenetic methods resulted in similar topologies with good support; the distal nodes were consistently weaker

whereas internal clades had an overall stronger support (Supplementary Figure 1, Chapter 2; Supplementary Figure 1, Chapter 3).

My analysis revealed that there were 3 CYP2 genes in the vertebrate ancestor; CYP2U, CYP2R and another CYP2 gene that lead to the diversification in vertebrates (Figure 1, Chapter 2). In the CYP4 analysis, likely single ancestral gene evolved into two clusters; one branch lead to the CYP4V and the invertebrate CYP4 cluster, and another evolved into a vertebrate cluster of CYP4 subfamilies (Figure 1, Chapter 3). Based on the CYP4 phylogeny, it is unlikely to find a basal CYP4 subfamily that would be found across all invertebrate and vertebrate species.

Taxonomic sampling in prior CYP2 and CYP4 studies was poor and included a limited selection of vertebrate species to illustrate the evolutionary relationship of the subfamilies. Avian and amphibian species were used as intermediate classes between the mammalian and actinopterygian classes, yet the numbers of sequences from these vertebrate classes were limited and placement in the phylogenetic tree not well resolved for some genes. This in part reflects the taxonomic sampling of genome projects, there were at least 4 mammalian and actinopterygiian species with completed genomes; fewer have been completed in the avian and amphibian classes. Amphibians are known for their developmental constraints, thus incorporating genomes from all 3 orders (anura, caudata, caecilians) would better represent the evolutionary diversity of amphibian species (2009). Fully sequenced avian genomes include chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*) genomes (Hackett *et al.* 2008). An increase in sequence availability in avian species would likely resolve poor topology placements of

avian CYP4B, CYP4F, and CYP2H genes. For example, including diverse species such as passerines, hummingbirds and loons from the largely diverse neognarthae infraclass and paleognathain species would enhance representation of avian evolution. Such limitations will be addressed by the vertebrate genome 10K sequencing (2009).

The diversity of fishes is extensive with over 50, 000 species including the jawless chondrichthyans, actinopterygians, and sarcopterygians (2009). Fishes are of extreme importance since they represent approximately 50% of vertebrates with an immense variation in morphology, physiology, ecology and present basic vertebrate developmental stages (2009). Including species from different classes (i.e. lamprey, shark, and lungfish) would illustrate the evolution of CYP families with higher resolution and likely strengthen support for placement of genes in CYP 2P/2AD/2N/2Z/2AE/2V and the CYP4V subfamilies. In the case of the CYP4V subfamilies, including the more distal lineages of fish would likely strengthen the placement of *Ciona* and sea urchin species.

### 4.2 Gene Annotation and Nomenclature

In vertebrate CYP families, gene structure may be a key element for identifying genes in a particular family and provides confidence in annotation of genomic sequences. CYP2R and CYP2U had 5 exons (Thomas 2007), whereas 9 exons were found in all other CYP2 subfamilies (Nelson *et al.* 2004) suggesting that the third ancestral vertebrate CYP2 gene had undergone intron insertions (Chapter 2). The gene structure of actinopterygian CYP2s was very similar to that found in mammalian sequences, providing greater confidence in the use of gene structure for annotation of CYP2s in non-mammalian lineages. Contrary to the CYP2s, exon number was more variable in the

CYP4 family and varied from 11 to 13, depending on the class and subfamilies of interest. Thus, gene structure is not as conserved in the CYP4 family and is not as informative for annotation of CYP4 genes in novel vertebrate genomes.

CYP nomenclature is typically based on sequence identity, families and subfamilies have amino acid sequence identity of >40% and >55%, respectively (Nelson *et al.* 2004). Phylogenetic analyses can be very important for the resolution of nomenclature for some genes. In my studies, a high resolution phylogenetic tree in many cases resolved or proposed new nomenclature. CYP genes from some public databases (e.g. Genbank) and early assemblies of genomes have a higher probability of incorrect nomenclature because the sequences have not necessarily been named by the CYP nomenclature committee. In my studies, some of these sequences clustered with different subfamilies, a strong indication that the nomenclature was incorrect. In other cases, the sequences may be too few to identify the most appropriate family or subfamily. For example, chicken CYP2C45 clustered with the CYP2H subfamily and amphibian CYP4T and avian CYP4B genes were not clustered monophyletically within their subfamilies. We have few avian and amphibian sequences and these classes may have unique CYP subfamilies that have yet to be identified. The *Ciona* and urchin sequences are problematic for nomenclature assignments, in part because they are evolutionarily distant to the large number of vertebrate sequences available for comparison. In addition, *Ciona* appear to have a high rate of substitutions (Goldstone *et al.* 2007). My phylogenetic trees identified nomenclature issues for several *Ciona* sequences. For example, the proposed CYP4F *Ciona* sequences clustered outside of the CYP4F subfamily, suggesting that they

likely belong to a totally different subfamily. Increased taxonomic sampling and robust phylogenetics will provide important information to identify incorrect nomenclature.

### 4.3 Subfamily expansions

CYP subfamily expansions are associated with recombination and tandem duplication events. In the case of the CYP2C subfamily, expansion was identified in humans and since they contain the L1 LINE repetitive sequences, a greater probability for recombination within this cluster exists. Greater CYP2C gene expansion was seen in rodents (15 genes in mice) than in humans (4 genes, Nelson *et al.* 2004). Of all vertebrate CYP families, only CYP families 2 and 4 had intermingled members in the same region of a chromosome (Nelson *et al.* 2004). Mirror symmetry was identified in the multi-loci clusters, CYP2ABFGS and CYP4ABZX, and are hypothesized to be driven by inverse and tandem duplication processes (Nelson *et al.* 2004). Interestingly, expansion of these subfamilies was much greater in rodents, although pseudogenes of the functional mouse genes were identified in the human genome (Nelson *et al.* 2004). For example, gene order was conserved in the CYP2 cluster between humans and rodents, with the exception that CYP2T was a pseudogene in humans (Hoffman *et al.* 2001). In vertebrate CYP families, tandem duplication events and inversion of genomic regions are strongly associated with largely expanded subfamilies or unique intermingling of subfamilies, in both cases these features seem to be quite distinct and have only been found in 7 clusters (Nelson *et al.* 2004).

## 4.4 Functional Divergence analyses

### *4.4.1 Type I Diverge analysis*

The functional understanding of the CYP superfamily has been primarily restricted to mammals with limited knowledge outside of this class of species. Interpreting functional divergence between the subfamilies can shed light on the functional similarities or differences, and this is best conducted within a phylogenetic context. The evolutionary functional divergence can be determined using DIVERGE 2.0 that calculates differences in evolutionary rates between pairwise groups (Gaucher *et al.* 2002). Regions with high theta values were thought to correlate with the SRS regions; yet results from Chapters 2 and 3 did not agree.

The evolutionary rates of functional divergence for CYP2 (Figure 4, Chapter 2) and CYP4 (Figure 2, Chapter3) analyses showed no particular clustering of high divergence. This is in contrast to a CYP3A study, where high divergence was identified and clustered within particular SRS regions (McArthur *et al.* 2003). Yet, small clusters of low divergence were found across subfamilies. Common regions of low divergence were found in the N-terminus region of the B-C helices, suggesting amino acid conservation yet, the region between helices B- C have been identified as part of the active site of the protein (Hasemann *et al.* 1995; Poulos *et al.* 1987). In both studies, higher divergence was associated with certain subfamilies, (Figure 4, Chapter 2; Figure 2, Chapter 3). CYP2A, CYP2E, and CYP4X subfamilies all had high evolutionary rates of

functional divergence in multiple pairwise comparisons covering the majority of the sequence alignment.

Pairwise comparisons that were not statistically significant based on the LRT values may be functionally similar. Some subfamilies with low pairwise divergence were expected; CYP2N and CYP2P were known to be functionally similar (Oleksiak *et al.* 2003). Yet, CYP2N and CYP2A subfamilies were not functionally divergent, which is surprising due to the large evolutionary distance between the subfamilies. Further investigation into the subfamilies with low functional divergence may reveal a plausible functional connection, thus enhancing the knowledge of orphan CYP genes in large families.

*4.4.2 Substrate Recognition Sites*

The CYP proteins were suggested to have six substrate recognition sites (SRSs) critical for substrate binding (Gotoh 1992). These regions were identified based on the high non-synonymous rates of substitution which aligned with the substrate interaction regions in the *P. putida* P450 101A (Gotoh 1992). These SRSs were determined using primarily mammalian CYP2 family sequences and were suggested to be functionally important for all CYPs (Gotoh 1992). The SRSs were not supported by the results from either CYP2 or CYP4 functional divergence analyses. In a study of the CYP3A subfamily (McArthur *et al.* 2003), some correlations were identified between regions with type I functional divergence and SRSs. It may be plausible to identify functionally divergent regions of CYP proteins at the subfamily level or within restricted taxonomic classes, yet this does not appear to hold true within a CYP family.

*4.4.3 Type II functional divergence*

Radical biochemical changes between closely related genes can be analyzed *in silico* using the type II functional divergence. This approach examines amino acids that are biochemically conserved in one group and yet different in another (ie. a hydrophobic vs. a hydrophilic amino acid; Gu 2006). The biochemical changes may be important for differences in substrate binding, metabolic rate or substrate preference. The focus in both studies (CYP2 and CYP4) was to identify radical biochemical changes in the active and substrate binding sites, to raise hypotheses regarding functional differences across sister subfamilies, across taxonomic classes within the same subfamily, or those specific residues that may be key for substrate specificity. Overall, statistically significant radical changes were in a range of $1<\theta< 8$, yet typical theta values ranged between 2 to 5.

The type II functionally divergent analyses that found a high number of significant residues also had the largest theta values and demonstrated clustering of radical sites within the active site. Comparisons with few changes showed low to no clustering. For example, the highest number of radical biochemical changes was in CYP2A/4F (Figure 7, Chapter 2) and CYP4X/4A (Figure 3, Chapter 3) pairwise comparisons, and clustering of radical residues was identified in active site regions in both analyses. Some sites with radical biochemical changes were found in the more conserved regions as well. In cases where type I analyses identified low or no divergence, the type II analyses almost always identified at least a few radical biochemically changed sites, these changes suggest that low rates of functional

divergence will have few radical biochemical changes (i.e. CYP4F22, Chapter 3 discussion; CYP2N/2P/2J, Chapter 2 discussion)

### 4.5 Digital Gene Expression (ESTs)

Gene expression is an integral part of determining function for putative genes. To generate a broad overview of CYP4 gene expression, especially in the non-mammalian species, a minimum of one species was selected from each vertebrate class for analyses. All mammalian libraries met the EST library cut-off limits, yet limited numbers and smaller library sizes were seen in the other 3 vertebrate classes. The inconsistency in library size and number across all species presents difficulty with conducting cross species comparisons. Those libraries that meet the minimum of 10, 000 ESTs are thought to represent a well sequenced tissue library, hence if no ESTs were detected that suggests very low or no expression in the tissue. For this approach to be most successful, an increase in library size and numbers outside of the mammalian class are needed, to increase the number of tissue libraries for more analogous comparisons.

This approach appears to be reasonable for examining differences in expression patterns. The expression patterns identified with EST libraries were in good agreement with published data from RT-PCR and Northern blots. Using EST data to identify expression patterns for homologous genes in different species could further functional hypothesis development and identify alternative tissue to target for gene expression. Strong CYP4X1 EST expression was found in the brain, trachea, vascular, and nerve tissue libraries that have an abundance of long chain fatty acid amides, and recently

CYP4X genes were hypothesized to function in ω-hydroxylation of long chain fatty acid amides based on their selective oxidation of anandamides in the brain.

### 4.6 Future Directions

Based on our phylogenetic analyses of the CYP2 family, it is difficult to infer functional specificity for non-mammalian subfamilies. *In silico* functional analyses (type I and II) did not suggest functional similarities between mammalian and non-mammalian subfamilies. Thus, determining function in non-mammalian species may require direct testing, one gene or subfamily at a time. High throughput screening of a CYP substrate library may be needed to identify function of the CYP2 genes. Characterizing the function of CYP2U and CYP2R in non-mammalian species will likely enhance the understanding of the CYP2 ancestral subfamilies. CYP2U1 function has not been widely determined, even in mammals, yet metabolic activity has been found with ω-hydroxylation of arachidonic acid (AA), decosahexaenoic acid, and long chain fatty acids (Chuang *et al.* 2004) providing substrates for functional testing of CYP2U homologs in other vertebrate species. Two AA metabolites, 19- and 20- HETE have important physiological functions (Chuang *et al.* 2004; Devos *et al.* 2010), which may also be important for non-mammalian vertebrates. Interestingly, type I functional divergence suggests functional similarity between the evolutionary distance subfamilies CYP2N and CYP2A, thus I suggest utilizing known CYP2A substrates for functional testing of CYP2N in fish. CYP2A are known to activate pro-carcinogens and drugs such as coumarin (Miles *et al.* 1990; Yamano *et al.* 1990; Yun *et al.* 1991), methoxyflurane

(Kharasch *et al.* 1995), and nicotine (Nakajima *et al.* 1996); these drugs may reveal a novel function of CYP2Ns in exogenous compound metabolism.

It would be of great interest to identify the functional importance of the mammalian CYP4F22 genes, based on the type I and II functional divergence results. Considering the low divergence of CYP4F22 with the other CYP4F genes, metabolism of eicosanoid and long chain fatty acids (Lefevre *et al.* 2006) may be expected. The potential for overlapping function across the CYP4F subfamily is suggested by my results. Lastly, functional overlap of CYP4X and CYP4F22 was suggested by type I functional divergence and could be the focus of specific functional testing. Since CYP4X proteins are capable of ω-hydroxylation of anandamides, identifying the catalytic potential of CYP4X proteins towards endogenous fatty acids, particularly long chain fatty acids, may aid the understanding of the physiological importance of this subfamily.

## 4.7 References

(2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100:659-74

Chuang SS, Helvig C, Taimi M, Ramshaw HA, Collop AH, Amad M, White JA, Petkovich M, Jones G, Korczak B (2004) CYP2U1, a novel human thymus- and brain-specific cytochrome P450, catalyzes omega- and (omega-1)-hydroxylation of fatty acids. J Biol Chem 279:6305-14

Devos A, Lino Cardenas CL, Glowacki F, Engels A, Lo-Guidice JM, Chevalier D, Allorge D, Broly F, Cauffiez C (2010) Genetic polymorphism of CYP2U1, a cytochrome P450 involved in fatty acids hydroxylation. Prostaglandins Leukot Essent Fatty Acids 83:105-10

Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315-21

Goldstone JV, Goldstone HM, Morrison AM, Tarrant A, Kern SE, Woodin BR, Stegeman JJ (2007) Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family. Mol Biol Evol 24:2619-31

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

Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23:1937-45

Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T (2008) A phylogenomic study of birds reveals their evolutionary history. Science 320:1763-8

Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure 3:41-62

Hoffman SM, Nelson DR, Keeney DS (2001) Organization, structure and evolution of the CYP2 gene cluster on human chromosome 19. Pharmacogenetics 11:687-98

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4

Kharasch ED, Hankins DC, Thummel KE (1995) Human kidney methoxyflurane and sevoflurane metabolism. Intrarenal fluoride production as a possible mechanism of methoxyflurane nephrotoxicity. Anesthesiology 82:689-99

Lefevre C, Bouadjar B, Ferrand V, Tadini G, Megarbane A, Lathrop M, Prud'homme JF, Fischer J (2006) Mutations in a new cytochrome P450 gene in lamellar ichthyosis type 3. Hum Mol Genet 15:767-76

McArthur AG, Hegelund T, Cox RL, Stegeman JJ, Liljenberg M, Olsson U, Sundberg P, Celander MC (2003) Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. J Mol Evol 57:200-11

Miles JS, McLaren AW, Forrester LM, Glancey MJ, Lang MA, Wolf CR (1990) Identification of the human liver cytochrome P-450 responsible for coumarin 7-hydroxylase activity. Biochem J 267:365-71

Nakajima M, Yamamoto T, Nunoya K, Yokoi T, Nagashima K, Inoue K, Funae Y, Shimada N, Kamataki T, Kuroiwa Y (1996) Role of human cytochrome P4502A6 in C-oxidation of nicotine. Drug Metab Dispos 24:1212-7

Nelson DR (1998) Metazoan cytochrome P450 evolution. Comp Biochem Physiol C Pharmacol Toxicol Endocrinol 121:15-22

Nelson DR (2003) Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys 409:18-24

Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14:1-18

Oleksiak MF, Wu S, Parker C, Qu W, Cox R, Zeldin DC, Stegeman JJ (2003) Identification and regulation of a new vertebrate cytochrome P450 subfamily, the CYP2Ps, and functional characterization of CYP2P3, a conserved arachidonic acid epoxygenase/19-hydroxylase. Arch Biochem Biophys 411:223-34

Poulos TL, Finzel BC, Howard AJ (1987) High-resolution crystal structure of cytochrome P450cam. J Mol Biol 195:687-700

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3:e67

Yamano S, Tatsuno J, Gonzalez FJ (1990) The CYP2A3 gene product catalyzes coumarin 7-hydroxylation in human liver microsomes. Biochemistry 29:1322-

9Yun CH, Shimada T, Guengerich FP (1991) Purification and characterization of human liver microsomal cytochrome P-450 2A6. Mol Pharmacol 40:679-85

# APPENDIX 1:

# SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 2

**Table A1.1. Comparison of the CYP2 subfamilies evolutionary divergence.** Assessment of the site-specific evolutionary rates of 17 vertebrate CYP2 subfamilies (with more than 4 gene representatives) via DIVERGE[a].

| CYP2 subfamilies | C | E | B | F | A | J | AD | N | P | Z | K | X | AA | Y | D | R | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C** | ~ | 0.364 + 0.092 | 0.422 ± 0.064 | 0.518 ± 0.090 | 0.232 ± 0.085 | 0.341 ± 0.041 | 0.370 ± 0.053 | 0.280 ± 0.062 | 0.383 ± 0.057 | 0.415 ± 0.068 | 0.366 ± 0.044 | 0.388 ± 0.057 | 0.453 ± 0.056 | 0.269 ± 0.064 | 0.306 ± 0.056 | 0.719 ± 0.062 | 0.488 ± 0.059 |
| **E** | 15.575 | ~ | 0.638± 0.110 | 0.836 ± 0.152 | 0.608 ± 0.157 | 0.614 ± 0.102 | 0.471 ± 0.128 | 0.622 ± 0.138 | 0.562 ± 0.151 | 0.674 ± 0.143 | 0.471 ± 0.092 | 0.379 ± 0.126 | 0.818 ± 0.133 | 0.731 ± 0.143 | 0.475 ± 0.131 | 0.759 ± 0.112 | 0.664 ± 0.120 |
| **B** | 42.807 | 33.630 | ~ | 0.718 ± 0.100 | 0.554 ± 0.117 | 0.413 ± 0.067 | 0.461 ± 0.093 | 0.453 ± 0.081 | 0.558 ± 0.076 | 0.580 ± 0.088 | 0.512 ± 0.059 | 0.496 ± 0.076 | 0.449 ± 0.079 | 0.313 ± 0.091 | 0.677 ± 0.099 | 0.633 ± 0.082 | 0.590 ± 0.083 |
| **F** | 32.819 | 30.060 | 51.305 | ~ | 0.642 ± 0.145 | 0.569 ± 0.099 | 0.326 ±0.139 | 0.333 ± 0.114 | 0.661 ± 0.107 | 0.437 ± 0.124 | 0.516 ± 0.082 | 0.460 ± 0.123 | 0.787 ± 0.118 | 0.465 ± 0.118 | 0.553 ± 0.120 | 0.530 ± .0.116 | 0.503 ± o.114 |
| **A** | 7.501 | 14.927 | 22.250 | 19.460 | ~ | 0.415 ± 0.086 | 0.658 ± 0.146 | 0.184 ± 0.159 | 0.634 ± 0.126 | 0.622 ± 0.127 | 0.573 ± 0.089 | 0.662 ± 0.118 | 0.612 ± 0.102 | 0.274 ± 0.120 | 0.726 ± 0.137 | 0.510 ± 0.111 | 0.649 ± 0.121 |
| **J** | 68.847 | 36.158 | 37.436 | 32.681 | 23.377 | ~ | 0.467 ± 0.069 | 0.301 ± 0.077 | 0.373 ± 0.064 | 0.453 ± 0.067 | 0.439 ± 0.051 | 0.569 ± 0.072 | 0.478 ± 0.058 | 0.313 ± 0.066 | 0.446 ± 0.076 | 0.449 ± 0.066 | 0.614 ± 0.075 |
| **AD** | 48.065 | 13.527 | 24.785 | 5.482 | 20.271 | 45.882 | ~ | 0.194 ± 0.086 | 0.371 ± 0.077 | 0.264 ± 0.074 | 0.411 ± 0.061 | 0.262 ± 0.074 | 0.432 ± 0.075 | 0.383 ± 0.098 | 0.293 ± 0.087 | 0.554 ± 0.080 | 0.420 ± 0.086 |
| **N** | 20.454 | 20.337 | 31.522 | 8.512 | 1.345 | 15.125 | 5.015 | ~ | 0.148 ± 0.080 | 0.155 ± 0.090 | 0.318 ± 0.058 | 0.272 ± 0.089 | 0.526 ± 0.084 | 0.116 ± 0.085 | 0.263 ± 0.107 | 0.446 ± 0.094 | 0.362 ± 0.097 |
| **P** | 44.810 | 13.901 | 53.216 | 37.948 | 25.466 | 33.917 | 23.309 | 3.391 | ~ | 0.322 ± 0.088 | 0.253 ± 0.048 | 0.298 ± 0.079 | 0.506 ± 0.080 | 0.464 ± 0.085 | 0.317 ± 0.101 | 0.554 ± 0.074 | 0.490 ± 0.077 |
| **Z** | 37.799 | 22.185 | 43.936 | 12.386 | 23.888 | 45.628 | 12.848 | 2.997 | 13.452 | ~ | 0.276 ± 0.055 | 0.266 ± 0.099 | 0.436 ± 0.069 | 0.344 ± 0.096 | 0.178 ± 0.099 | 0.631 ± 0.087 | 0.378 ± 0.084 |
| **K** | 68.448 | 26.347 | 74.049 | 39.270 | 41.322 | 73.912 | 45.176 | 29.763 | 27.866 | 24.821 | ~ | 0.330 ± 0.051 | 0.466 ± 0.053 | 0.320 ± 0.069 | 0.290 ± 0.070 | 0.510 ± 0.054 | 0.614 ± 0.065 |
| **X** | 47.184 | 9.063 | 42.397 | 13.967 | 31.425 | 61.698 | 12.543 | 9.283 | 14.245 | 7.293 | 40.963 | ~ | 0.462 ± 0.073 | 0.177 ± 0.079 | 0.346 ± 0.098 | 0.571 ± 0.077 | 0.533 ± 0.087 |
| **AA** | 65.105 | 52.858 | 31.983 | 44.542 | 35.671 | 68.819 | 33.021 | 39.253 | 39.637 | 40.107 | 78.849 | 39.819 | ~ | 0.336 ± 0.084 | 0.414 ± 0.076 | 0.838 ± 0.074 | 0.705 ± 0.076 |
| **Y** | 17.722 | 26.244 | 11.834 | 15.601 | 5.215 | 22.130 | 15.171 | 1.861 | 29.722 | 12.941 | 21.702 | 4.961* | 15.964 | ~ | 0.486 ± 0.103 | 0.469 ± 0.090 | 0.482 ± 0.089 |
| **D** | 29.711 | 13.228 | 46.381 | 21.206 | 28.044 | 34.888 | 11.385 | 6.045 | 9.819 | 3.275 | 17.071 | 12.540 | 30.121 | 22.175 | ~ | 0.602 ± 0.088 | 0.546 ± 0.104 |
| **R** | 134.674 | 45.576 | 59.533 | 20.809 | 21.097 | 46.015 | 47.846 | 22.747 | 55.402 | 52.856 | 90.432 | 54.992 | 127.026 | 27.245 | 47.268 | ~ | 0.428 ± 0.077 |
| **U** | 67.719 | 30.370 | 50.313 | 19.417 | 28.723 | 67.249 | 23.590 | 14.020 | 40.132 | 20.301 | 88.160 | 37.157 | 85.847 | 29.429 | 27.637 | 30.563 | ~ |

[a] Measured theta ($\vartheta$), coefficient of evolutionary functional divergence, and their respective standard error are shown above the diagonal. An increasing of functional divergence values are represented by $\theta \geq 0.5$ up to 1. Below the diagonal are the likelihood ratio test (LRT) values , to test the null hypothesis of $\theta = 0$. Highlighted LRT values significantly reject the null hypothesis ($p \leq 0.05$)

**Table A1.2.** CYP2 sequences included in the alignment and their associated retrieval sources (de novo, P450 Homepage, and accession numbers from NCBI).

| Labeled sequence in the alignment | Gene retrieval source | Accession # | Chromosome | Start bp | Stop bp | EST/ Scaffold |
|---|---|---|---|---|---|---|
| Cow CYP2A13 | *P450 Homepage* | | | | | CB434432.1, CB463229, TC193989 |
| Cow CYP2B6 | *de novo + P450 Homepage* | NC_007316.4 | 18 | 49962063 | 49945946 | |
| Cow CYP2C85 | *de novo + P450 Homepage* | NC_007327.4 | 26 | 16760713 | 16788513 | |
| Cow CYP2C87 | *de novo + P450 Homepage* | AC_000183.1 | 26 | 16030318 | 16065061 | |
| Cow CYP2C88 | *de novo + P450 Homepage* | AC_000183.1 | 26 | 16231685 | 16288072 | |
| Cow CYP2C89 | *de novo + P450 Homepage* | NC_007327.4 | 26 | 17007986 | 16960998 | |
| Cow CYP2C90 | *de novo + P450 Homepage* | NC_007327.4 | 26 | 16559058 | 16601508 | |
| Cow CYP2D14 | *de novo + P450 Homepage* | NC_007303.4 | 5 | 120047146 | 120043003 | |
| Cow CYP2D43 | *de novo + P450 Homepage* | NC_007303.4 | 5 | 120034713 | 120030551 | |
| Cow CYP2E1 | *de novo + P450 Homepage* | NC_007327.4 | 26 | 5856553 | 5866827 | |
| Cow CYP2F1 | *de novo + P450 Homepage* | NC_007316.4 | 18 | 49729060 | 49719599 | |
| Cow CYP2G1 | *de novo + P450 Homepage* | NC_007316.4 | 18 | 49827401 | 49814773 | |
| Cow CYP2J26 | *de novo + P450 Homepage* | AC_000160.1 | 3 | 86490320 | 86519377 | |
| Cow CYP2J27 | *de novo + P450 Homepage* | AC_000160.1 | 3 | 86526816 | 86547282 | |
| Cow CYP2J28 | *de novo + P450 Homepage* | NC_007301.4 | 3 | 92254381 | 92294882 | |
| | | | | | | SCAFFOLD25620 DN825584.1, DN825689.1, CN793001.1, SCAFFOLD126970 SCAFFOLD87184 CK957631.1 SCAFFOLD166503 CK834395 CK945908.1 SCAFFOLD166503 CK960783.1 CK957063.1 SCAFFOLD245478 CK945645.1 |
| Cow CYP2J29 | *P450 Homepage* | | | | | |
| Cow CYP2J30 | *de novo + P450 Homepage* | NC_007301.4 | 3 | 92411328 | 92380138 | |
| Cow CYP2R1 | *de novo + P450 Homepage* | ENSCowG00000010419 | 15 | 36674395 | 36698051 | |
| Cow CYP2S1 | *de novo + P450 Homepage* | NC_007316.4 | 18 | 50031837 | 50043866 | |
| Cow CYP2U1 | *de novo + P450 Homepage* | NC_007304.4 | 6 | 18684795 | 18666546 | |
| Dog CYP2A13 CYP2A6[1] | *de novo* | NW_876270.1_NW_876270.1 | 1 | 43235496_43229491 | 43228866_43235490 | |
| Dog CYP2A7 | *de novo* | NW_876270.1 | 1 | 43197750 | 43203984 | |
| Dog CYP2B | *de novo* | NW_876270.1 | 1 | 43114807 | 43129385 | |
| Dog CYP2C | *de novo* | NW_876285.1 | 28 | 8748107 | 8725176 | |
| Dog CYP2E | *de novo* | NW_876287.1 | 28 | 395882 | 405665 | |
| Dog CYP2F | *de novo* | NW_876270.1 | 1 | 43272128 | 43283098 | |
| Dog CYP2J | *de novo* | NW_876313.1 | 5 | 19914056 | 19956047 | |
| Dog CYP2R | *de novo* | NW_876273.1 | 21 | 37769697 | 37744500 | |
| Dog CYP2S | *de novo* | NW_876270.1 | 1 | 43044442 | 43033920 | |
| Dog CYP2U | *de novo* | ENSCAFG00000011203 | 32 | 31348146 | 31366257 | |
| Dog CYP2W | *de novo* | NW_876319.1 | 6 | 293563 | 287849 | |
| Ciona CYP2J30 | NCBI | XP 002124426 | | | | |
| Ciona CYP2N | NCBI | XP 002125554 | | | | |
| Ciona CYP2U1 | NCBI | XP 002124096 | | | | |
| Zebrafish CYP2AA1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA10 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA11 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA12 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA2 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA3v2 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA4 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA6 | *P450 Homepage* | | | | | |

| Labeled sequence in the alignment | Gene retrieval source | Accession # | Chromosome | Start bp | Stop bp | EST/ Scaffold |
|---|---|---|---|---|---|---|
| Zebrafish CYP2AA7 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA8 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA9v1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AA9v2 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AD2 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AD3 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AD6 | *P450 Homepage* | | | | | |
| Zebrafish CYP2AE1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K16 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K17 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K19 (NP 001073172.1) | *P450 Homepage* | | | | | |
| Zebrafish CYP2K19 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K20 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K21 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K6 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K7 | *P450 Homepage* | | | | | |
| Zebrafish CYP2K8 | *P450 Homepage* | | | | | |
| Zebrafish CYP2N13 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P10 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P12 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P6 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P7 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P8 | *P450 Homepage* | | | | | |
| Zebrafish CYP2P9 | *P450 Homepage* | | | | | |
| Zebrafish CYP2R1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2U1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2V1 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X10 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X11 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X12 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X6 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X7 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X8 | *P450 Homepage* | | | | | |
| Zebrafish CYP2X9 | *P450 Homepage* | | | | | |
| Zebrafish CYP2Y3 | *P450 Homepage* | | | | | |
| Zebrafish CYP2Y4 | *P450 Homepage* | | | | | |
| Fugu CYP2K10 | *P450 Homepage* | | | | | |
| Fugu CYP2K11 | *P450 Homepage* | | | | | |
| Fugu CYP2K9 | *P450 Homepage* | | | | | |
| Fugu CYP2N10 | *P450 Homepage* | | | | | |
| Fugu CYP2N11 | *P450 Homepage* | | | | | |
| Fugu CYP2N12 | *P450 Homepage* | | | | | |
| Fugu CYP2N9 | *P450 Homepage* | | | | | |
| Fugu CYP2R1 | *P450 Homepage* | | | | | |
| Fugu CYP2U1 | *P450 Homepage* | | | | | |
| Fugu CYP2X2 | *P450 Homepage* | | | | | |
| Fugu CYP2Y1 | *P450 Homepage* | | | | | |
| Fugu CYP2Y2 | *P450 Homepage* | | | | | |
| Fugu CYP2Z1 | *P450 Homepage* | | | | | |
| Fugu CYP2Z2 | *P450 Homepage* | | | | | |
| Stickleback CYP2K23 | *de novo + Ensembl* | | 11 | 9797706 | 9794340 | |
| Stickleback CYP2K24 | *de novo + Ensembl* | | 11 | 9720130 | 9723292 | |
| Stickleback CYP2K25 | *de novo + Ensembl* | | 11 | 9679866 | 9676172 | |
| Stickleback CYP2K26 | *de novo + Ensembl* | | 18 | 12864956 | 12862311 | DN708008.1 |
| Stickleback CYP2N15 | *de novo + Ensembl* | | 8 | 19111310 | 19114905 | CD506195.1, CD504080.1, CD507761.1 |

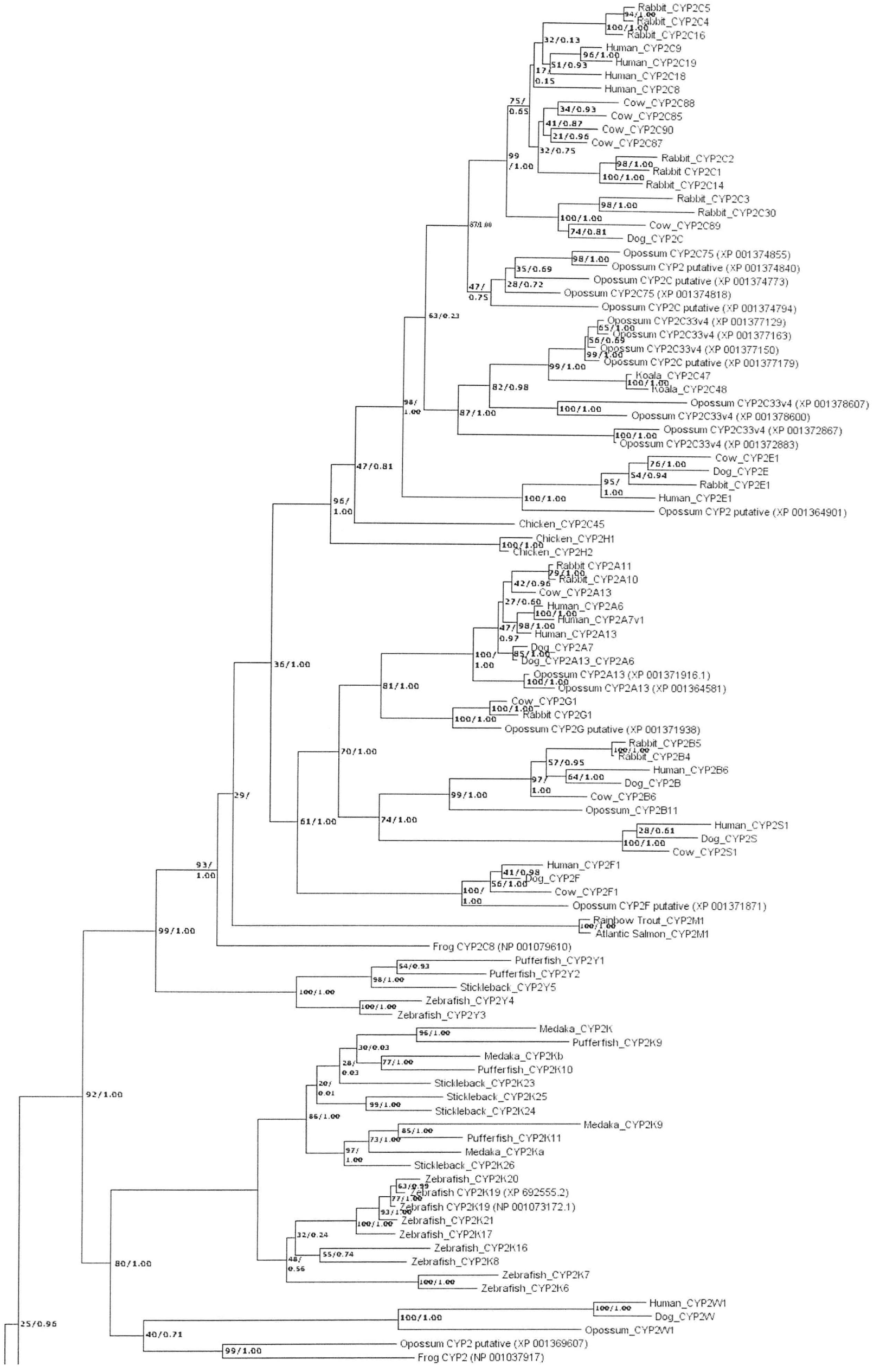
| Labeled sequence in the alignment | Gene retrieval source | Accession # | Chromosome | Start bp | Stop bp | EST/ Scaffold |
|---|---|---|---|---|---|---|
| Stickleback CYP2N16 | *de novo* + *Ensembl* | | 8 | 19116077 | 19119925 | |
| Stickleback CYP2N17 | *de novo* + *Ensembl* | | 16 | 2232906 | 2228494 | DT966028.1, DW631570.1, |
| Stickleback CYP2R1 | *de novo* + *Ensembl* | | 2 | 9716096 | 8718824 | |
| Stickleback CYP2U1 | *de novo* + *Ensembl* | | 9 | 8022276 | 8019743 | |
| Stickleback CYP2X13 | *de novo* + *Ensembl* | | 19 | 19948531 | 19940205 | |
| Stickleback CYP2Y5 | *de novo* + *Ensembl* | | 1 | 16592713 | 16588688 | |
| Stickleback CYP2Z3 | *de novo* + *Ensembl* | | 8 | 15172608 | 15168449 | |
| Stickleback CYP2Z4 | *de novo* + *Ensembl* | | 8 | 15161857 | 15158571 | |
| Chicken CYP2C45 | NCBI + *P450 Homepage* | NP 001001752 | | | | |
| Chicken CYP2H1 | NCBI + *P450 Homepage* | NP 001001616 | | | | |
| Chicken CYP2H2 | P450 Homepage | | | | | |
| Chicken CYP2J2 | NCBI + *P450 Homepage* | XP 422553 | | | | |
| Chicken CYP2J20 | P450 Homepage | | | | | |
| Chicken CYP2J21 | NCBI + *P450 Hon epage* | CAG32696 | | | | |
| Chicken CYP2J23 | NCBI + *P450 Homepage* | XP 422509 | | | | |
| Chicken CYP2R1 | NCBI | XP 420996 | | | | |
| Human CYP2A13 | P450 Homepage | | | | | |
| Human CYP2A6 | P450 Homepage | | | | | |
| Human CYP2A7v1 | P450 Homepage | | | | | |
| Human CYP2B6 | P450 Homepage | | | | | |
| Human CYP2C18 | P450 Homepage | | | | | |
| Human CYP2C19 | P450 Homepage | | | | | |
| Human CYP2C8 | P450 Homepage | | | | | |
| Human CYP2C9 | P450 Homepage | | | | | |
| Human CYP2D6 | P450 Homepage | | | | | |
| Human CYP2E1 | P450 Homepage | | | | | |
| Human CYP2F1 | P450 Homepage | | | | | |
| Human CYP2J2 | P450 Homepage | | | | | |
| Human CYP2R1 | P450 Homepage | | | | | |
| Human CYP2S1 | P450 Homepage | | | | | |
| Human CYP2U1 | P450 Homepage | | | | | |
| Human CYP2W1 | P450 Homepage | | | | | |
| Opossum CYP2 (XP 001364901) | NCBI | XP 001364901 | | | | |
| Opossum CYP2 (XP 001374840) | NCBI | XP 001374840 | | | | |
| Opossum CYP2 putative (XP 001369607) | NCBI | XP 001369607 | | | | |
| Opossum CYP2C putative (XP 001374773) | NCBI | XP 001374773 | | | | |
| Opossum CYP2C putative (XP 001374794) | NCBI | XP 001374794 | | | | |
| Opossum CYP2C putative (XP 001377179) | NCBI | XP 001377179 | | | | |
| Opossum CYP2C33v4 (XP 001372867; | NCBI | XP 001372867 | | | | |
| Opossum CYP2C33v4 (XP 001372883) | NCBI | XP 001372883 | | | | |
| Opossum CYP2C33v4 (XP 001377129) | NCBI | XP 001377129 | | | | |
| Opossum CYP2C33v4 (XP 001377150) | NCBI | XP 001377150 | | | | |
| Opossum CYP2C33v4 (XP 001377163) | NCBI | XP 001377163 | | | | |
| Opossum CYP2C33v4 (XP 001378600) | NCBI | XP 001378600 | | | | |
| Opossum CYP2C33v4 (XP 001378607) | NCBI | XP 001378607 | | | | |
| Opossum CYP2C75 (XP 001374818) | NCBI | XP 001374818 | | | | |
| Opossum CYP2C75 (XP 001374855) | NCBI | XP 001374855 | | | | |
| Opossum CYP2F putative (XP 001371871) | NCBI | XP 001371871 | | | | |
| Opossum CYP2G putative (XP 001371938) | NCBI | XP 001371938 | | | | |

| Labeled sequence in the alignment | Gene retrieval source | Accession # | Chromosome | Start bp | Stop bp | EST/ Scaffold |
|---|---|---|---|---|---|---|
| Opossum CYP2J putative (XP 001381096) | NCBI | XP 001381096 | | | | |
| Opossum CYP2J2 (XP 001381050) | NCBI | XP 001381050 | | | | |
| Opossum CYP2J2 (XP 001381099) | NCBI | XP 001381099 | | | | |
| Opossum CYP2J2 (XP 001381107) | NCBI | XP 001381107 | | | | |
| Opossum CYP2U putative (XP 001366063) | NCBI | XP 001366063 | | | | |
| Opossum CYP2A13 (XP 001371916.1) | NCBI | XP 001371916.1 | | | | |
| Opossum CYP2A13 (XP 001364581) | NCBI | XP 001364581 | | | | |
| Opossum CYP2B11 | NCBI | XP 001371957 | | | | |
| Opossum CYP2R1 | NCBI | XP 001378967 | | | | |
| Opossum CYP2W1 | NCBI | XP 001378319 | | | | |
| Platypus CYP2R1 | NCBI | XP 001518489 | | | | |
| Rabbit CYP2A10 | NCBI | AAA31371 | | | | |
| Rabbit CYP2A11 (L10237.1) | NCBI | AAA31372 | | | | |
| Rabbit CYP2B4 | NCBI | P00178 | | | | |
| Rabbit CYP2B5 | NCBI | P12789 | | | | |
| Rabbit CYP2C1(P00180) | NCBI | P00180 | | | | |
| Rabbit CYP2C14 | NCBI | P17666 | | | | |
| Rabbit CYP2C16 | NCBI | AAA31227 | | | | |
| Rabbit CYP2C2 | NCBI | P00181 | | | | |
| Rabbit CYP2C3 | NCBI | P00182 | | | | |
| Rabbit CYP2C30 | NCBI | BAA05140 | | | | |
| Rabbit CYP2C4 | NCBI | P11371 | | | | |
| Rabbit CYP2C5 | NCBI | AAA63461 | | | | |
| Rabbit CYP2D23 | NCBI | BAA84472 | | | | |
| Rabbit CYP2G1 (P24461) | NCBI | P24461 | | | | |
| Rabbit CYP2D24 | NCBI | BAA84473 | | | | |
| Rabbit CYP2E1 | NCBI | P08682 | | | | |
| Rabbit CYP2J1 | NCBI | BAA14401 | | | | |
| MedakaCYP2AD (de novo) | *de novo* + *Ensembl* | | 4 | 28094594 | 28090211 | |
| MedakaCYP2K (de novo) | *de novo* + *Ensembl* | | 8 | 11125163 | 11122047 | |
| MedakaCYP2Ka(de novo) | *de novo* + *Ensembl* | | 24 | 11284466 | 11289272 | |
| Medaka CYP2Kb | *de novo* + *Ensembl* | | 8 | 11131969 | 11128887 | |
| MedakaCYP2K9 (de novo) | *de novo* + *Ensembl* | | 24 | 11296473 | 11301307 | |
| MedakaCYP2Na (de novo) | *de novo* + *Ensembl* | | 4 | 28073969 | 28070480 | |
| MedakaCYP2Nb (de novo) | *de novo* + *Ensembl* | | 4 | 28087146 | 28082821 | |
| MedakaCYP2P (de novo) | *de novo* + *Ensembl* | | 4 | 2813085 | 28120258 | |
| MedakaCYP2P3 (ABQ44505) | NCBI | ABQ44505 | | | | |
| MedakaCYP2R (de novo) | *de novo* + *Ensembl* | | 3 | 17796105 | 17801798 | |
| MedakaCYP2U (de novo) | *de novo* + *Ensembl* | | 1 | 20317219 | 20324274 | |
| MedakaCYP2X (de novo) | *de novo* + *Ensembl* | | 6 | 21437485 | 21427961 | |
| MedakaCYP2Z (de novo) | *de novo* + *Ensembl* | | 4 | 31518512 | 31523602 | |
| Rainbow trout CYP2M1 | NCBI | Q92088 | | | | |
| Koala CYP2C47 | NCBI | ACB98739 | | | | |
| Koala CYP2C48 | NCBI | ACB98740 | | | | |
| Atlantic Salmon CYP2M1 | NCBI | ACI33595 | | | | |
| Frog CYP2 (NP 001037917); | NCBI | NP 001037917 | | | | |
| Frog CYP2C8 (NP 001079610) | NCBI | NP 001079610 | | | | |
| Frog CYP2D6 (NP 001015719) | NCBI | NP 001015719 | | | | |
| Frog CYP2D6 (NP 001079789) | NCBI | NP 001079789 | | | | |
| Frog CYP2D6 (NP 001087043) | NCBI | NP 001087043 | | | | |

1 Dog_CYP2A13 and Dog_CYP2A6 have very high sequence identify and formed one phylogenetic branch, therefore; only one sequence was used throughout the study.

## Appendix Figures

**Figure A.1.1. Vertebrate CYP2 phylogenetic analysis**. Detailed phylogenetic analysis of the CYP2 proteins in vertebrates; a total of 24 CYP2 subfamilies were included. A JTT+I+G (Jones Taylor Thornton + invariant sites + gamma distribution) model was selected using ProtTest (Abascal *et al.* 2005). A maximum likelihood approach (RAxML) (Stamatakis *et al.* 2008) was applied to the final alignment of 196 sequences to determine the phylogenetic history. Bayesian inference was computed based on the final CYP2 alignment. All nodes are labeled with bootstrap values and posterior probabilities. Accession numbers are provided for opossum sequences, as these lack full gene nomenclature. Accession numbers for all other sequences are provided in Supplementary Table 2.


Rabbit_CYP2C5
Rabbit_CYP2C4
Rabbit_CYP2C16
Human_CYP2C9
Human_CYP2C19
Human_CYP2C18
Human_CYP2C8
Cow_CYP2C88
Cow_CYP2C85
Cow_CYP2C90
Cow_CYP2C87
Rabbit_CYP2C2
Rabbit CYP2C1
Rabbit_CYP2C14
Rabbit_CYP2C3
Rabbit_CYP2C30
Cow_CYP2C89
Dog_CYP2C
Opossum CYP2C75 (XP 001374855)
Opossum CYP2 putative (XP 001374840)
Opossum CYP2C putative (XP 001374773)
Opossum CYP2C75 (XP 001374818)
Opossum CYP2C putative (XP 001374794)
Opossum CYP2C33v4 (XP 001377129)
Opossum CYP2C33v4 (XP 001377163)
Opossum CYP2C33v4 (XP 001377150)
Opossum CYP2C putative (XP 001377179)
Koala_CYP2C47
Koala_CYP2C48
Opossum CYP2C33v4 (XP 001378607)
Opossum CYP2C33v4 (XP 001378600)
Opossum CYP2C33v4 (XP 001372867)
Opossum CYP2C33v4 (XP 001372883)
Cow_CYP2E1
Dog_CYP2E
Rabbit_CYP2E1
Human_CYP2E1
Opossum CYP2 putative (XP 001364901)
Chicken_CYP2C45
Chicken_CYP2H1
Chicken_CYP2H2
Rabbit CYP2A11
Rabbit_CYP2A10
Cow_CYP2A13
Human_CYP2A6
Human_CYP2A7v1
Human_CYP2A13
Dog_CYP2A7
Dog_CYP2A13_CYP2A6
Opossum CYP2A13 (XP 001371916.1)
Opossum CYP2A13 (XP 001364581)
Cow_CYP2G1
Rabbit CYP2G1
Opossum CYP2G putative (XP 001371938)
Rabbit_CYP2B5
Rabbit_CYP2B4
Human_CYP2B6
Dog_CYP2B
Cow_CYP2B6
Opossum_CYP2B11
Human_CYP2S1
Dog_CYP2S
Cow_CYP2S1
Human_CYP2F1
Dog_CYP2F
Cow_CYP2F1
Opossum CYP2F putative (XP 001371871)
Rainbow Trout_CYP2M1
Atlantic Salmon_CYP2M1
Frog CYP2C8 (NP 001079610)
Pufferfish_CYP2Y1
Pufferfish_CYP2Y2
Stickleback_CYP2Y5
Zebrafish_CYP2Y4
Zebrafish_CYP2Y3
Medaka_CYP2K
Pufferfish_CYP2K9
Medaka_CYP2Kb
Pufferfish_CYP2K10
Stickleback_CYP2K23
Stickleback_CYP2K25
Stickleback_CYP2K24
Medaka_CYP2K9
Pufferfish_CYP2K11
Medaka_CYP2Ka
Stickleback_CYP2K26
Zebrafish_CYP2K20
Zebrafish CYP2K19 (XP 692555.2)
Zebrafish CYP2K19 (NP 001073172.1)
Zebrafish_CYP2K21
Zebrafish_CYP2K17
Zebrafish_CYP2K16
Zebrafish_CYP2K8
Zebrafish_CYP2K7
Zebrafish_CYP2K6
Human_CYP2W1
Dog_CYP2W
Opossum_CYP2W1
Opossum CYP2 putative (XP 001369607)
Frog CYP2 (NP 001037917)

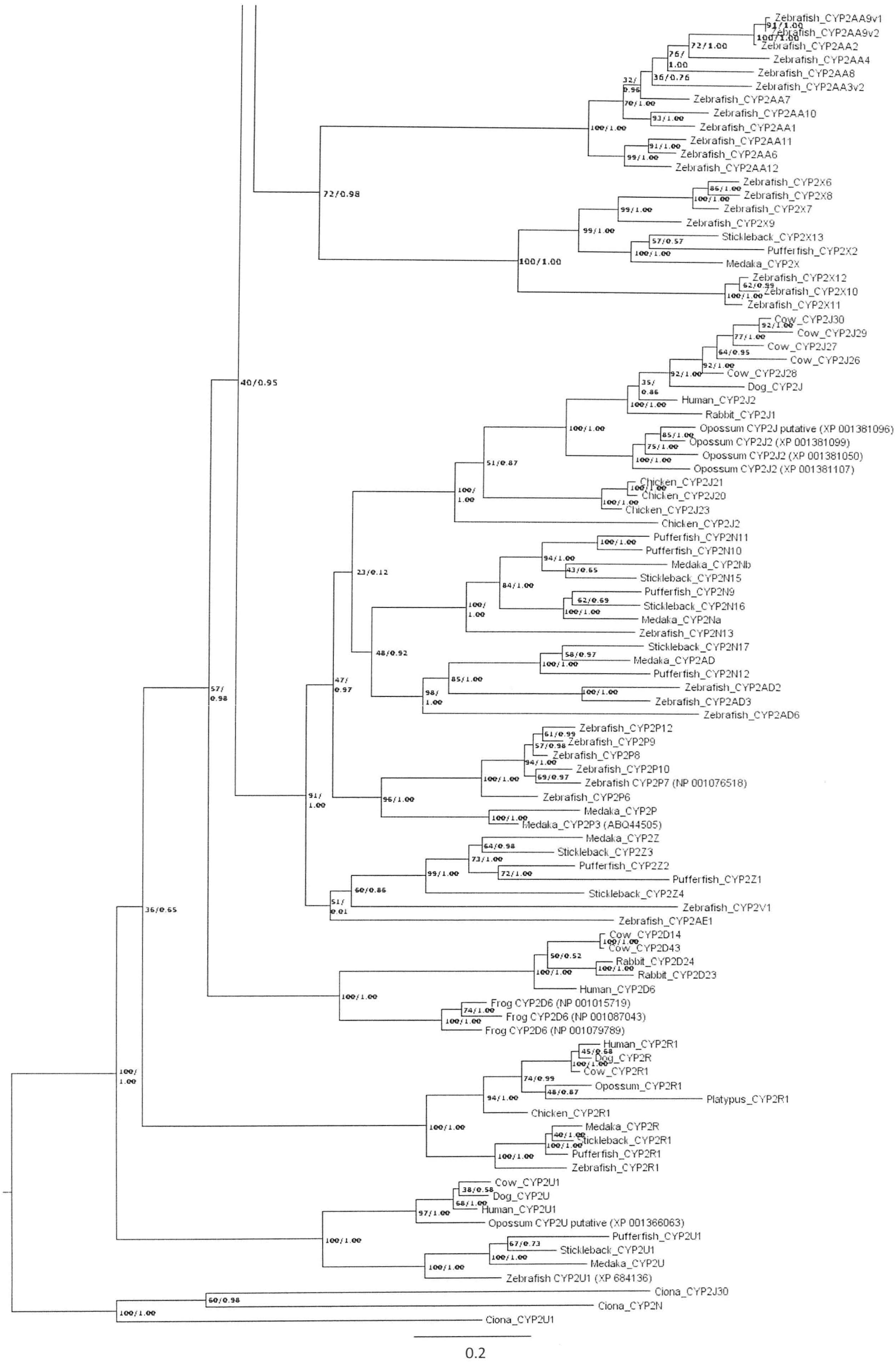

Zebrafish_CYP2AA9v1
Zebrafish_CYP2AA9v2
Zebrafish_CYP2AA2
Zebrafish_CYP2AA4
Zebrafish_CYP2AA8
Zebrafish_CYP2AA3v2
Zebrafish_CYP2AA7
Zebrafish_CYP2AA10
Zebrafish_CYP2AA1
Zebrafish_CYP2AA11
Zebrafish_CYP2AA6
Zebrafish_CYP2AA12
Zebrafish_CYP2X6
Zebrafish_CYP2X8
Zebrafish_CYP2X7
Zebrafish_CYP2X9
Stickleback_CYP2X13
Pufferfish_CYP2X2
Medaka_CYP2X
Zebrafish_CYP2X12
Zebrafish_CYP2X10
Zebrafish_CYP2X11
Cow_CYP2J30
Cow_CYP2J29
Cow_CYP2J27
Cow_CYP2J26
Cow_CYP2J28
Dog_CYP2J
Human_CYP2J2
Rabbit_CYP2J1
Opossum CYP2J putative (XP 001381096)
Opossum CYP2J2 (XP 001381099)
Opossum CYP2J2 (XP 001381050)
Opossum CYP2J2 (XP 001381107)
Chicken_CYP2J21
Chicken_CYP2J20
Chicken_CYP2J23
Chicken_CYP2J2
Pufferfish_CYP2N11
Pufferfish_CYP2N10
Medaka_CYP2Nb
Stickleback_CYP2N15
Pufferfish_CYP2N9
Stickleback_CYP2N16
Medaka_CYP2Na
Zebrafish_CYP2N13
Stickleback_CYP2N17
Medaka_CYP2AD
Pufferfish_CYP2N12
Zebrafish_CYP2AD2
Zebrafish_CYP2AD3
Zebrafish_CYP2AD6
Zebrafish_CYP2P12
Zebrafish_CYP2P9
Zebrafish_CYP2P8
Zebrafish_CYP2P10
Zebrafish CYP2P7 (NP 001076518)
Zebrafish_CYP2P6
Medaka_CYP2P
Medaka_CYP2P3 (ABQ44505)
Medaka_CYP2Z
Stickleback_CYP2Z3
Pufferfish_CYP2Z2
Pufferfish_CYP2Z1
Stickleback_CYP2Z4
Zebrafish_CYP2V1
Zebrafish_CYP2AE1
Cow_CYP2D14
Cow_CYP2D43
Rabbit_CYP2D24
Rabbit_CYP2D23
Human_CYP2D6
Frog CYP2D6 (NP 001015719)
Frog CYP2D6 (NP 001087043)
Frog CYP2D6 (NP 001079789)
Human_CYP2R1
Dog_CYP2R
Cow_CYP2R1
Opossum_CYP2R1
Platypus_CYP2R1
Chicken_CYP2R1
Medaka_CYP2R
Stickleback_CYP2R1
Pufferfish_CYP2R1
Zebrafish_CYP2R1
Cow_CYP2U1
Dog_CYP2U
Human_CYP2U1
Opossum CYP2U putative (XP 001366063)
Pufferfish_CYP2U1
Stickleback_CYP2U1
Medaka_CYP2U
Zebrafish CYP2U1 (XP 684136)
Ciona_CYP2J30
Ciona_CYP2N
Ciona_CYP2U1
0.2

**Figure A.1.2. Phylogenetic analysis of the CYP2J / 2P/ 2N / 2AD / 2AE / 2V / 2Z cluster**. A phylogenetic analysis of the CYP2J-2Z cluster (node 7 in Figure 1) based on a maximum likelihood approach is shown. Bootstrap values and posterior probabilities are labeled on all nodes. The CYP2 alignment included 343 amino acids for this analysis. *Ciona* CYP2N, CYP2J and CYP2X sequences were used as outgroups. Accession numbers are provided for opossum sequences, as these lack full gene nomenclature. Accession numbers for all other sequences are provided in Supplementary Table 2.

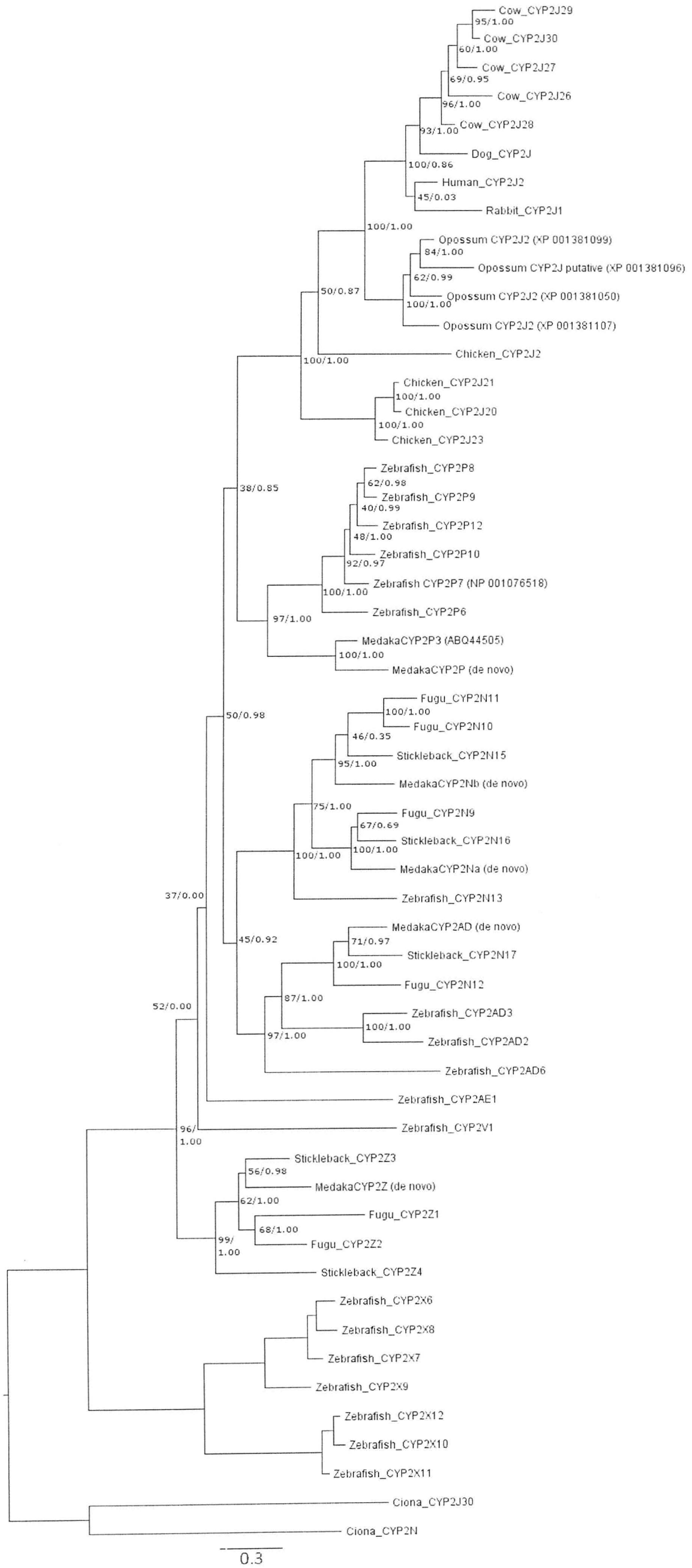

Cow_CYP2J29
95/1.00
Cow_CYP2J30
60/1.00
Cow_CYP2J27
69/0.95
Cow_CYP2J26
96/1.00
Cow_CYP2J28
93/1.00
Dog_CYP2J
100/0.86
Human_CYP2J2
45/0.03
Rabbit_CYP2J1
100/1.00
Opossum CYP2J2 (XP 001381099)
84/1.00
Opossum CYP2J putative (XP 001381096)
62/0.99
Opossum CYP2J2 (XP 001381050)
50/0.87
100/1.00
Opossum CYP2J2 (XP 001381107)
Chicken_CYP2J2
100/1.00
Chicken_CYP2J21
100/1.00
Chicken_CYP2J20
100/1.00
Chicken_CYP2J23
Zebrafish_CYP2P8
62/0.98
Zebrafish_CYP2P9
40/0.99
Zebrafish_CYP2P12
38/0.85
48/1.00
Zebrafish_CYP2P10
92/0.97
Zebrafish CYP2P7 (NP 001076518)
100/1.00
Zebrafish_CYP2P6
97/1.00
MedakaCYP2P3 (ABQ44505)
100/1.00
MedakaCYP2P (de novo)
Fugu_CYP2N11
100/1.00
Fugu_CYP2N10
50/0.98
46/0.35
Stickleback_CYP2N15
95/1.00
MedakaCYP2Nb (de novo)
75/1.00
Fugu_CYP2N9
67/0.69
Stickleback_CYP2N16
100/1.00
100/1.00
MedakaCYP2Na (de novo)
37/0.00
Zebrafish_CYP2N13
MedakaCYP2AD (de novo)
71/0.97
45/0.92
Stickleback_CYP2N17
100/1.00
Fugu_CYP2N12
87/1.00
52/0.00
Zebrafish_CYP2AD3
100/1.00
97/1.00
Zebrafish_CYP2AD2
Zebrafish_CYP2AD6
Zebrafish_CYP2AE1
96/
1.00
Zebrafish_CYP2V1
Stickleback_CYP2Z3
56/0.98
MedakaCYP2Z (de novo)
62/1.00
Fugu_CYP2Z1
68/1.00
99/
1.00
Fugu_CYP2Z2
Stickleback_CYP2Z4
Zebrafish_CYP2X6
Zebrafish_CYP2X8
Zebrafish_CYP2X7
Zebrafish_CYP2X9
Zebrafish_CYP2X12
Zebrafish_CYP2X10
Zebrafish_CYP2X11
Ciona_CYP2J30
Ciona_CYP2N
0.3

**Figure A.1.3. Phylogenetic analysis of the CYP2C clade.** A phylogenetic analysis of the CYP2C cluster (node 2 in Figure 1) based on a maximum likelihood approach is shown. Bootstrap values and posterior probabilities are labeled on all nodes. The CYP2 alignment included 346 amino acids for this analysis. The outgroup sequences were from the CYP2E (mammalian) subfamily. Accession numbers are provided for opossum sequences, as these lack full gene nomenclature. Accession numbers for all other sequences are provided in Supplementary Table 2.

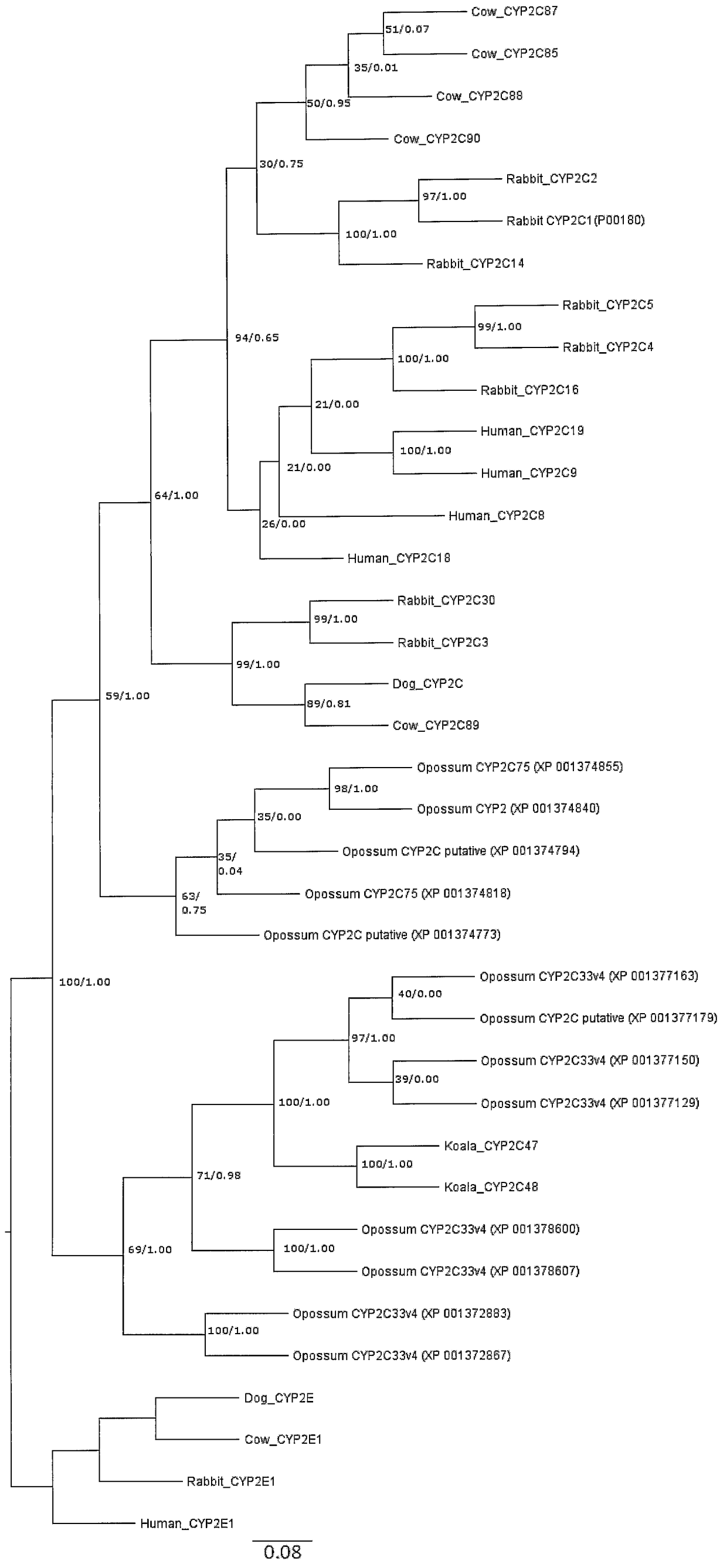

Cow_CYP2C87
51/0.07
Cow_CYP2C85
35/0.01
Cow_CYP2C88
50/0.95
Cow_CYP2C90
30/0.75
Rabbit_CYP2C2
97/1.00
Rabbit CYP2C1 (P00180)
100/1.00
Rabbit_CYP2C14
Rabbit_CYP2C5
99/1.00
Rabbit_CYP2C4
94/0.65
100/1.00
Rabbit_CYP2C16
21/0.00
Human_CYP2C19
100/1.00
21/0.00
Human_CYP2C9
64/1.00
Human_CYP2C8
26/0.00
Human_CYP2C18
Rabbit_CYP2C30
99/1.00
Rabbit_CYP2C3
99/1.00
Dog_CYP2C
59/1.00
89/0.81
Cow_CYP2C89
Opossum CYP2C75 (XP 001374855)
98/1.00
Opossum CYP2 (XP 001374840)
35/0.00
Opossum CYP2C putative (XP 001374794)
35/
0.04
Opossum CYP2C75 (XP 001374818)
63/
0.75
Opossum CYP2C putative (XP 001374773)
100/1.00
Opossum CYP2C33v4 (XP 001377163)
40/0.00
Opossum CYP2C putative (XP 001377179)
97/1.00
Opossum CYP2C33v4 (XP 001377150)
39/0.00
100/1.00
Opossum CYP2C33v4 (XP 001377129)
Koala_CYP2C47
100/1.00
71/0.98
Koala_CYP2C48
Opossum CYP2C33v4 (XP 001378600)
69/1.00
100/1.00
Opossum CYP2C33v4 (XP 001378607)
Opossum CYP2C33v4 (XP 001372883)
100/1.00
Opossum CYP2C33v4 (XP 001372867)
Dog_CYP2E
Cow_CYP2E1
Rabbit_CYP2E1
Human_CYP2E1
0.08

**Figure A.1.4. Speciation patterns and gene duplications of CYP 2W, 2X, 2AA and 2K subfamilies.** The evolution of vertebrate subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ★ ). The two CYP2 putative sequences (Opossum CYP2 putative (XP_001369607) and Frog CYP2 (NP_001037917) are strongly supported through phylogenetic analysis and are closely clustered with CYP2W, a mammalian subfamily.

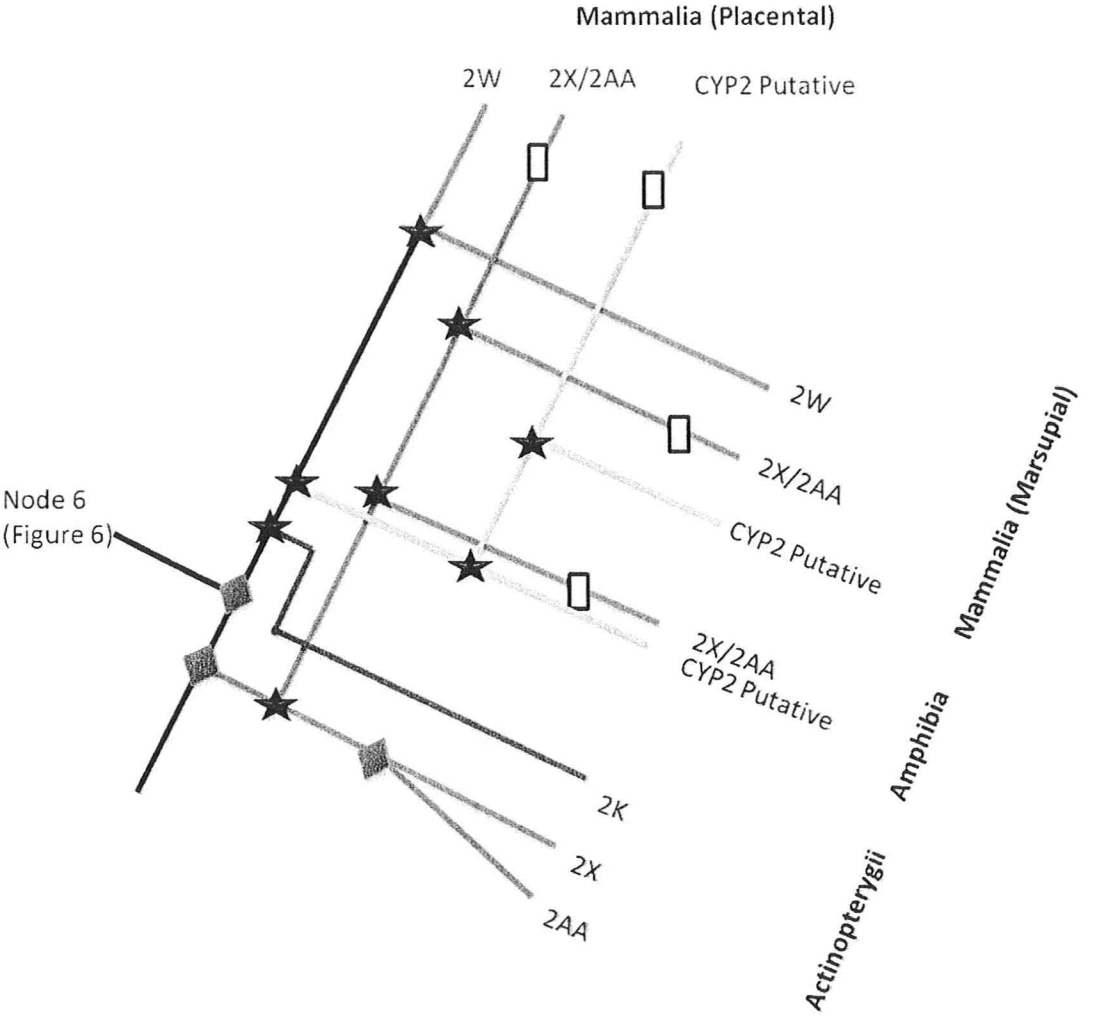

Mammalia (Placental)
2W
2X/2AA
CYP2 Putative
2W
2X/2AA
CYP2 Putative
2X/2AA
CYP2 Putative
Node 6
(Figure 6)
2K
2X
2AA
Mammalia (Marsupial)
Amphibia
Actinopterygii

**Figure A.1.5. Speciation patterns and gene duplications of CYP 2A, 2Y, 2F, 2B, 2S and 2G subfamilies.** The evolution of vertebrate subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ★ ). Dashed lines represent areas of topology with weak support. Frog 2C8 is not a true 2C sequence, it does not cluster with the 2C sequences. Frog 2C8 sequence is clustered between CYP2Y and CYP2M, both are actinopterygian subfamilies.

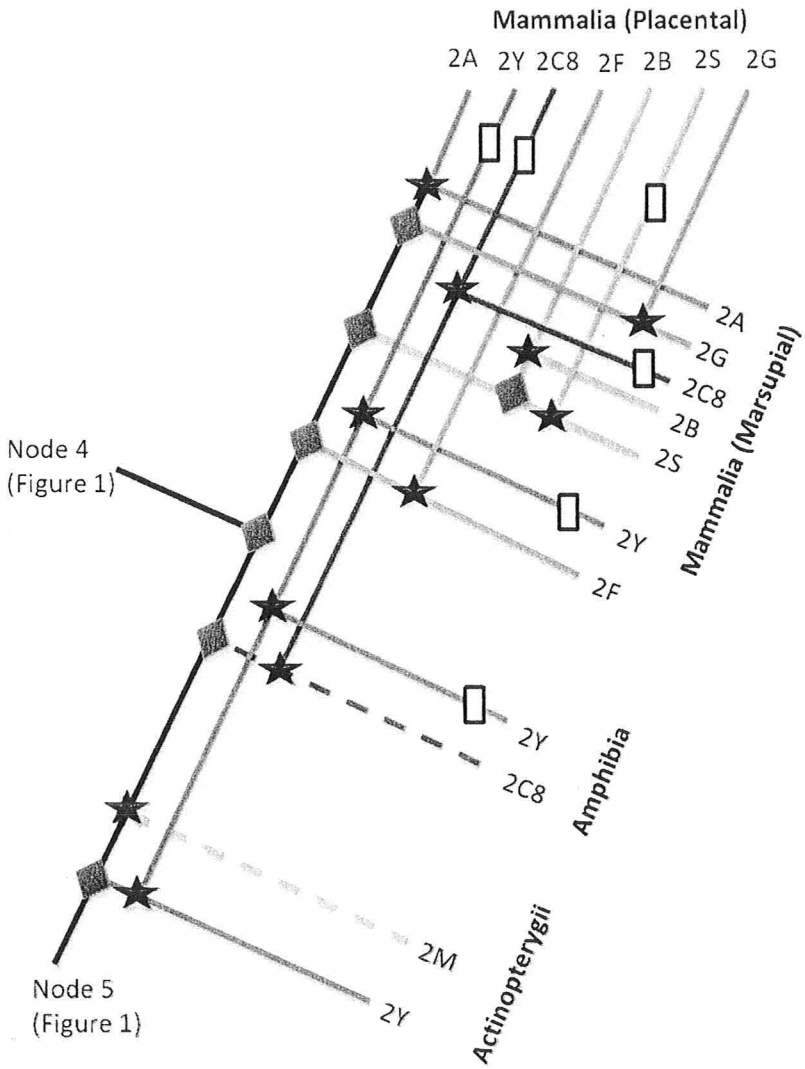
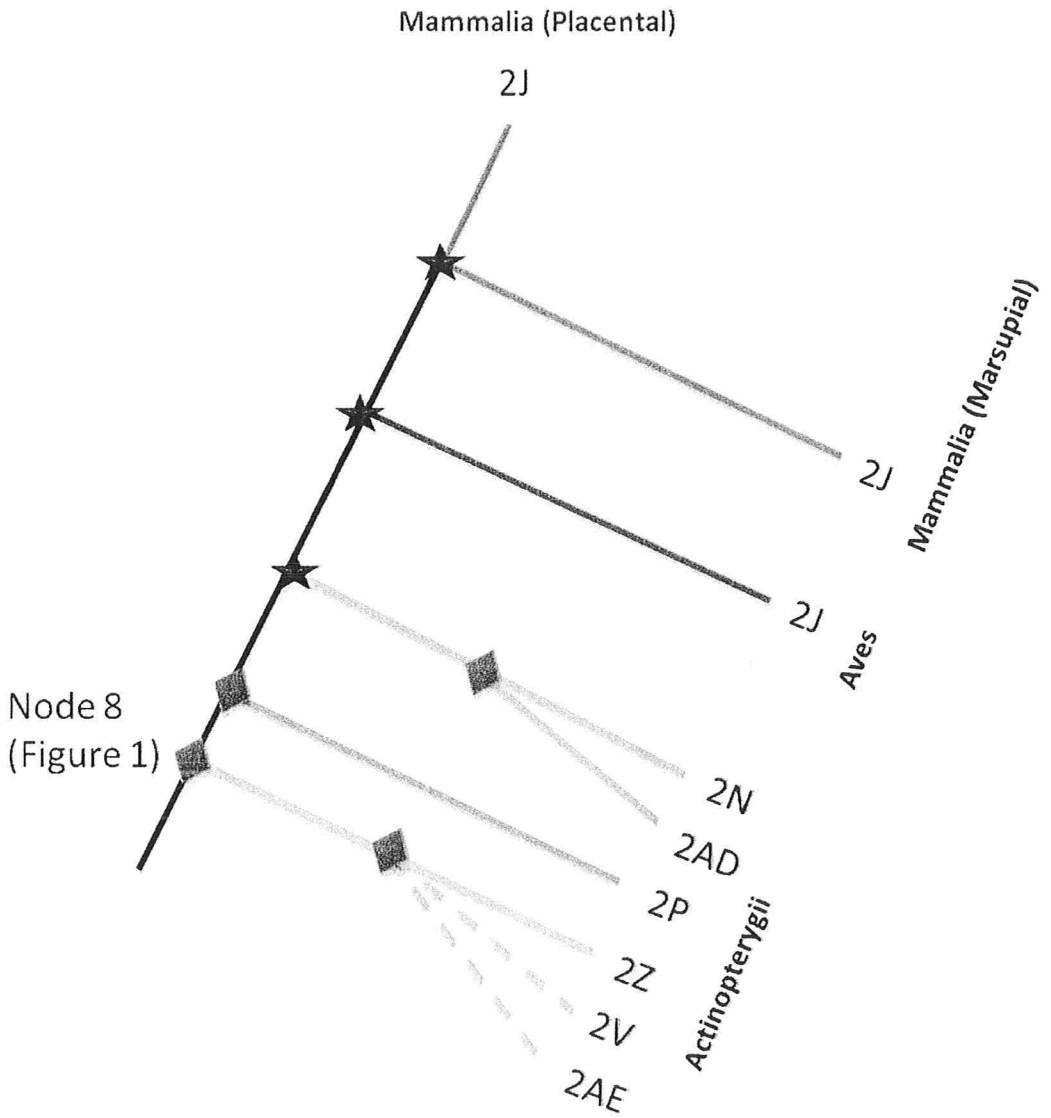

Mammalia (Placental)
2A 2Y 2C8 2F 2B 2S 2G
2A
2G
2C8
2B
2S
2Y
2F
Mammalia (Marsupial)
Node 4
(Figure 1)
2Y
2C8
Amphibia
2M
2Y
Actinopterygii
Node 5
(Figure 1)

**Figure A.1.6. Speciation patterns and gene duplications of CYP 2J, 2N, 2AD, 2P, 2Z, 2V and 2AE subfamilies.** The evolution of vertebrate subfamilies are symbolized by duplication ( ) and gene loss ( ) events. Diversification of vertebrate species is symbolized by speciation patterns ( ). Dashed lines represent areas of topology with weak support. This subfamily cluster presented with complexity in its topology as seen with poor support for the CYP2V and CYP2AE subfamilies which are zebrafish specific. The speciation and duplication patterns were derived from the original CYP2 phylogenetic analysis (Supplementary Figure 1) and not from distinct analysis (Supplementary Figure 2).


Mammalia (Placental)
2J
2J
Mammalia (Marsupial)
2J
Aves
Node 8
(Figure 1)
2N
2AD
2P
2Z
2V
2AE
Actinopterygii

# APPENDIX 2:

# SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 3

**Table A2.1. CYP4 and outgroup (CYP46) sequences.** The full names of sequences used in the alignment (CYP4.nex) and phylogenetic tree (Supplementary figure 1) are provided along with the accession numbers , where available. Those sequences that were *de novo* annotations are noted.

| Sequences used in the alignment (CYP4_aln.nex) | Sequence within the CYP4 tree | Accession numbers | *De novo* additions |
|---|---|---|---|
| **Mammalia** | | | |
| Cow_CYP46a_gi\|116003839\|ref\|NP | Cow_CYP46a_NP_001070278 | NP_001070278 | |
| Cow_CYP4A11_514aa_gi\|118150926 | Cow_CYP4A11_NP_001071376 | NP_001071376 | *De novo* additions |
| Cow_CYP4A22_515aa_gi\|149642723 | Cow_CYP4A22_NP_001092460 | NP_001092460 | *De novo* additions |
| Cow_CYP4B_511aa_gi\|115495419\|r | Cow_CYP4B_NP_001069670 | NP_001069670 | *De novo* additions |
| Cow_CYP4F2a_523aa_gi\|119894605 | Cow_CYP4F2a_XP_586481 | XP_586481 | *De novo* additions |
| Cow_CYP4F2b_524aa_gi\|78042534\| | Cow_CYP4F2b_NP_001030214 | NP_001030214 | *De novo* additions |
| Cow_CYP4F2c_523aa_gi\|115496710 | Cow_CYP4F2c_NP_001068790 | NP_001068790 | *De novo* additions |
| Cow_CYP4F3_522aa_gi\|114052010\| | Cow_CYP4F3_NP_001039856 | NP_001039856 | *De novo* additions |
| Cow_CYP4V_527aa_gi\|77735695\|re | Cow_CYP4V_NP_001029545 | NP_001029545 | *De novo* additions |
| Cow_CYP4X_515aa_gi\|194665816\|r | Cow_CYP4X_XP_592978 | XP_592978 | *De novo* additions |
| Dog_CYP4A37_510aa_99%_gi\|114158646 | Dog_CYP4A37_NP_001041490 | NP_001041490 | *De novo* additions |
| Dog_CYP4A38_511aa_100%_gi\|114158640 | Dog_CYP4A38_NP_001041482 | NP_001041482 | *De novo* additions |
| Dog_CYP4A39_511aa_100%_gi\|114158644 | Dog_CYP4A39_NP_001041499 | NP_001041499 | *De novo* additions |
| Dog_CYP4B1_511aa_12exons_100%_73977060 | Dog_CYP4B1_XP_850244 | XP_850244 | *De novo* additions |
| Dog_CYP4F22_532aa_ENSCAFT00000025291 | Dog_CYP4F22_ENSCAFT00000025291 | ENSCAFT00000025291 | *De novo* additions |
| Dog_CYP4F6_gi\|73986390\|ref\|XP_541 | Dog_CYP4F6_XP_541983 | XP_541983 | *De novo* additions |
| Dog_CYP4V2_Newer_version_73979556 | Dog_CYP4V2_XP_849364 | XP_849364 | *De novo* additions |
| Dog_CYP4X1_507aa_gi\|73977808 | Dog_CYP4X1_XP_539622 | XP_539622 | *De novo* additions |
| Giant Panda_\|281343177\|ref | Giant Panda_EFB18761 | EFB18761 | |
| Giant Panda_\|281349859\| | Giant Panda_EFB25443 | EFB25443 | |
| Goat_CYP4B2_\|25246586\|ref | Goat_CYP4B2_AAN72309 | AAN72309 | |
| Horse_CYP4A11_\|194207489\|ref\|XP_ | Horse_CYP4A11_XP_001495099 | XP_001495099 | |
| Horse_CYP4B1_\|149694470\|ref\|XP_0 | Horse_CYP4B1_XP_001495287 | XP_001495287 | |
| Horse_CYP4B1_\|194207491\|ref\|XP_0 | Horse_CYP4B1_XP_001495202 | XP_001495202 | |
| Horse_CYP4F18_\|149759130\|ref\|XP_ | Horse_CYP4F18_XP_001500868 | XP_001500868 | |
| Horse_CYP4F22_\|194223692\|ref\|XP_ | Horse_CYP4F22_XP_001914769 | XP_001914769 | |
| Horse_CYP4F3a_\|194223697\|ref\|XP_ | Horse_CYP4F3a_XP_001914864 | XP_001914864 | |
| Horse_CYP4F3b_\|194238911\|ref\|XP_ | Horse_CYP4F3b_XP_001915272 | XP_001915272 | |
| Horse_CYP4X1_\|194207487\|ref\|XP_0 | Horse_CYP4X1_XP_001493525 | XP_001493525 | |
| Human_CYP46 | Human_CYP46 | * | |
| Human_CYP4A11_520aa_gi\|4503235\|r | Human_CYP4A11_NP_001073 | NP_001073 | |
| Human_CYP4A22_520aa_gi\|62952506\| | Human_CYP4A22_NP_001010969 | NP_001010969 | |
| Human_CYP4B1_512aa_gi\|153218660\| | Human_CYP4B1_NP_000770 | NP_000770 | |
| Human_CYP4F_521aa_gi\|13435391\|re | Human_CYP4F_NP_001073 | NP_001073 | |
| Human_CYP4F11_gi\|193083178\|ref\|N | Human_CYP4F11_NP_067010 | NP_067010 | |
| Human_CYP4F12_525aa_gi\|150036266 | Human_CYP4F12_NP_076433 | NP_076433 | |

| Sequences used in the alignment (CYP4_aln.nex) | Sequence within the CYP4 tree | Accession numbers | *De novo* additions |
|---|---|---|---|
| Human_CYP4F3_521aa_gi\|119220562\| | Human_CYP4F3_NP_000887 | NP_000887 | |
| Human_CYP4F8_521aa_gi\|6005737\|re | Human_CYP4F8_NP_009184 | NP_009184 | |
| Human_CYP4V2_526aa_gi\|187960086\| | Human_CYP4V2_NP_997235 | NP_997235 | |
| Human_CYP4X1_510aa_gi\|29837648\|r | Human_CYP4X1_NP_828847 | NP_828847 | |
| Human_CYP4Z1_506aa_gi\|30023836\|r | Human_CYP4Z1_NP_835235 | NP_835235 | |
| Koala_CYP4A15_\|27462772\|ref | Koala_CYP4A15_AAO15579 | AAO15579 | |
| Minke Whale_CYP4A35_\|149689500\|ref | Minke Whale_CYP4A35_BAF64511 | BAF64511 | |
| Minke Whale_CYP4V6_\|149689502\|ref | Minke Whale_CYP4V6_BAF64512 | BAF64512 | |
| Opossum_CYP4F11_\|126324002\|ref\|XP_ | Opossum_CYP4F11_XP_001365295 | XP_001365295 | |
| Opossum_CYP4F3a_\|126324071\|ref\|XP_ | Opossum_CYP4F3a_XP_001367758 | XP_001367758 | |
| Opossum_CYP4F3b_\|126324079\|ref\|XP_ | Opossum_CYP4F3b_XP_001367837 | XP_001367837 | |
| Opossum_CYP4F3c_\|126324083\|ref\|XP_ | Opossum_CYP4F3c_XP_001367881 | XP_001367881 | |
| Opossum_CYP4V2_\|126331227\|ref\|XP_0 | Opossum_CYP4V2_XP_001368368 | XP_001368368 | |
| Opossum_Pred_\|126324075\|ref | Opossum_XP_001367797 | XP_001367797 | |
| Pig_CYP4_\|47523904\|ref\|NP_9995 | Pig_CYP4_NP_999590 | NP_999590 | |
| Pig_CYP4A24_\|47523902\|ref | Pig_CYP4A24_NP_999589 | NP_999589 | |
| Platypus_CYP4B1_\|149609226\|ref | Platypus_CYP4B1_XP_001519517 | XP_001519517 | |
| Platypus_Pred_\|149577062\|ref | Platypus_XP_001520882 | XP_001520882 | |
| Rabbit_CYP4A4_\|50403715\|sp\|P10611 | Rabbit_CYP4A4_P10611 | P10611 | |
| Rabbit_CYP4A5_\|283436163\|ref | Rabbit_CYP4A5_NP_001164448 | NP_001164448 | |
| Rabbit_CYP4A6_\|283806689\|ref | Rabbit_CYP4A6_NP_001164599 | NP_001164599 | |
| Rabbit_CYP4A7_\|283806687\|ref | Rabbit_CYP4A7_NP_001164598 | NP_001164598 | |
| Rabbit_CYP4F2_\|291413170\|ref\|XP_0 | Rabbit_CYP4F2_XP_002722843 | XP_002722843 | |
| Rabbit_CYP4F3a_\|291410589\|ref\|XP_ | Rabbit_CYP4F3a_XP_002721578 | XP_002721578 | |
| Rabbit_CYP4F3b_\|291413160\|ref\|XP_ | Rabbit_CYP4F3b_XP_002722849 | XP_002722849 | |
| Rabbit_CYP4F3c_\|291413162\|ref\|XP_ | Rabbit_CYP4F3c_XP_002722841 | XP_002722841 | |
| Rabbit_CYP4F3d_\|291413164\|ref\|XP_ | Rabbit_CYP4F3d_XP_002722842 | XP_002722842 | |
| Rabbit_CYP4F3e_\|291413168\|ref\|XP_ | Rabbit_CYP4F3e_XP_002722851 | XP_002722851 | |
| Rabbit_CYP4V2_\|291386017\|ref\|XP_0 | Rabbit_CYP4V2_XP_002709379 | XP_002709379 | |
| Rabbit_CYP4X1_\|291398956\|ref\|XP_0 | Rabbit_CYP4X1_XP_002715705 | XP_002715705 | |
| Sheep_CYP4F21_\|57619220\|ref | Sheep_CYP4F21_NP_001009743 | NP_001009743 | |
| **Aves** | | | |
| Chicken_4A/4B | Chicken_putative_CYP4A/4B | * | |
| Chicken_4V_531aa_gi\|50657412\|ref\|N | Chicken_CYP4V_NP_001001879 | NP_001001879* | |
| Chicken_CYP4F_514aa_New_family | Chicken_CYP4F | * | |
| Zebra finch_CYP4B1a_\|224058125\|ref\|XP_ | Zebra finch_CYP4B1a_XP_002192090 | XP_002192090 | |
| Zebra finch_CYP4B1b_\|224058127\|ref\|XP_ | Zebra finch_CYP4B1b_XP_002192124 | XP_002192124 | |

| Sequences used in the alignment (CYP4_aln.nex) | Sequence within the CYP4 tree | Accession numbers | *De novo* additions |
|---|---|---|---|
| **Amphibia** | | | |
| Frog_CYP4_515aa_gi\|147899821\|re | Frog_CYP4_NP_001079027 | NP_001079027 | |
| Frog_CYP46_gi\|112418634\|gb\|AAI2 | Frog_CYP46_AAI22063 | AAI22063 | |
| Frog_CYP46_gi\|118404892\|ref\|NP_ | Frog_CYP46_NP_001072550 | NP_001072550 | |
| Frog_CYP46A_gi\|187608754\|ref\|NP | Frog_CYP46A_NP_001120275 | NP_001120275 | |
| Frog_CYP4B_496aa_gi\|147905750\|r | Frog_CYP4B_NP_001090538 | NP_001090538 | |
| Frog_CYP4B_510aa_gi\|148225196\|r | Frog_CYP4B_NP_001090539 | NP_001090539 | |
| Frog_CYP4B_510aa_gi\|187608807\|r | Frog_CYP4B_NP_001120073 | NP_001120073 | |
| Frog_CYP4B_515aa_gi\|62860148\|re | Frog_CYP4B_NP_001017348 | NP_001017348 | |
| Frog_CYP4F2_528aa_gi\|288557252\| | Frog_CYP4F2_NP_001165650 | NP_001165650 | |
| Frog_CYP4F22_528aa_gi\|62751474\| | Frog_CYP4F22_NP_001015810 | NP_001015810 | |
| Frog_CYP4F22_529aa_gi\|148231945 | Frog_CYP4F22_NP_001091388 | NP_001091388 | |
| Frog_CYP4V2__522aa_gi\|148229743 | Frog_CYP4V2__NP_001086053 | NP_001086053 | |
| Frog_CYP4V2_523aa_gi\|118404542\| | Frog_CYP4V2_NP_001072667 | NP_001072667 | |
| Frog_CYP4V4_520aa_gi\|288557254\| | Frog_CYP4V4_NP_001165651 | NP_001165651 | |
| **Actinopterygii** | | | |
| Atlantic salmon_CYP4F3_\|209154704\|ref | Atlantic salmon_CYP4F3_ACI33584 | ACI33584 | |
| European seabass_CYP_\|4091078\|ref | European seabass_CYP_AAC98961 | * | |
| Fugu_CYP4F28_ENSTRUT00000001144 | Fugu_CYP4F28_ENSTRUT00000001144 | ENSTRUT00000001144 | *De novo* additions |
| Fugu_CYP4T_ENSTRUT0000000819 | Fugu_CYP4T_ENSTRUT0000000819 | ENSTRUT0000000819 | *De novo* additions |
| Fugu_CYP4V5_ENSTRUT00000010988 | Fugu_CYP4V5_ENSTRUT00000010988 | ENSTRUT00000010988 | *De novo* additions |
| Green pufferfish_\|47216297\|ref | Green pufferfish_CAF96593 | CAF96593 | |
| Medaka_CYP4B_ENSORLT00000017909_74%_ | Medaka_CYP4T_ENSORLT00000017909 | ENSORLT00000017909 | *De novo* additions |
| Medaka_CYP4F_ENSORLT00000008103_1 | Medaka_CYP4F_ENSORLT00000008103 | ENSORLT00000008103 | *De novo* additions |
| Medaka_CYP4V_ENSORLT00000010255_78%_ | Medaka_CYP4V_ENSORLT00000010255 | ENSORLT00000010255 | *De novo* additions |
| Stickleback_CYP4F_putative_ENSGACT00000019452 | Stickleback_CYP4F_putative_ENSGACT00000019452 | ENSGACT00000019452 | *De novo* additions |
| Stickleback_CYP4T_putative_ENSGACT00000018457 | Stickleback_CYP4T_putative_ENSGACT00000018457 | ENSGACT00000018457 | *De novo* additions |
| Stickleback_CYP4V_putative_ENSGACP00000022294 | Stickleback_CYP4V_putative_ENSGACP00000022294 | ENSGACP00000022294 | *De novo* additions |
| Zebrafish_CYP46a | Zebrafish_CYP46a | * | |
| Zebrafish_CYP46b | Zebrafish_CYP46b | * | |
| Zebrafish_CYP4F_gi\|148230579\|ref\|NP_ | Zebrafish_CYP4F_NP_001083010 | * | |
| Zebrafish_CYP4T_\|40363539\|ref\|NP_954 | Zebrafish_CYP4T_NP_954686 | * | |
| Zebrafish_CYP4Va_gi\|117606212\|ref\|NP | Zebrafish_CYP4Va_NP_001071070 | NP_001071070 | |
| Zebrafish_CYP4Vb_gi\|121583883\|ref\|NP | Zebrafish_CYP4Vb_NP_001073465 | NP_001073465 | |
| **Ascidiacea** | | | |
| Ciona_CYP4F17_\|198419762\| | Ciona_CYP4F17_XP_002130606 | XP_002130606 | |
| Ciona_CYP4F22_\|198424634\| | Ciona_CYP4F22_XP_002129693 | XP_002129693 | |
| Ciona_CYP4F40_\|198435280\| | Ciona_CYP4F40_XP_002132033 | XP_002132033 | |
| Ciona_CYP4V2_\|198436000\| | Ciona_CYP4V2_XP_002132121 | XP_002132121 | |

| Sequences used in the alignment (CYP4_aln.nex) | Sequence within the CYP4 tree | Accession numbers | *De novo* additions |
|---|---|---|---|
| Sea Urchin_]115625653\|ref\|XP_783176.2\| | Sea Urchin_XP_783176 | XP_783176 | |
| Sea Urchin_]115647018\|ref\|XP_784930.2\| | Sea Urchin_XP_784930 | XP_784930 | |
| Sea Urchin_]72014091\|ref\|XP_786946.1\| | Sea Urchin_XP_786946 | XP_786946 | |
| Sea Urchin_gi\|115625651\|ref\|XP_783244.2\| | Sea Urchin_XP_783244 | XP_783244 | |
| **Gastropoda** | | | |
| Snail_CYP4_]189092908\|gb\|ACD7582 | Snail_CYP4_ACD75824 | ACD75824 | |
| Snail_CYP4_]189092910\|gb\|ACD7582 | Snail_CYP4_ACD75825 | ACD75825 | |
| Snail_CYP4_]189092914\|gb\|ACD7582 | Snail_CYP4_ACD75827 | ACD75827 | |
| Snail_CYP4_]189092920\|gb\|ACD7583 | Snail_CYP4_ACD75830 | ACD75830 | |
| **Insecta** | | | |
| Drosophila_CYP4AA1 | Drosophila_CYP4AA2 | * | |
| Drosophila_CYP4AC1 | Drosophila_CYP4AC1 | * | |
| Drosophila_CYP4AC2 | Drosophila_CYP4AC2 | * | |
| Drosophila_CYP4AC3 | Drosophila_CYP4AC3 | * | |
| Drosophila_CYP4AD1 | Drosophila_CYP4AD2 | * | |
| Drosophila_CYP4AE1 | Drosophila_CYP4AE2 | * | |
| Drosophila_CYP4C3 | Drosophila_CYP4C4 | * | |
| Drosophila_CYP4D1 | Drosophila_CYP4D1 | * | |
| Drosophila_CYP4D14 | Drosophila_CYP4D14 | * | |
| Drosophila_CYP4D2 | Drosophila_CYP4D2 | * | |
| Drosophila_CYP4D20 | Drosophila_CYP4D20 | * | |
| Drosophila_CYP4D21 | Drosophila_CYP4D21 | * | |
| Drosophila_CYP4D8 | Drosophila_CYP4D8 | * | |
| Drosophila_CYP4E1 | Drosophila_CYP4E1 | * | |
| Drosophila_CYP4E2 | Drosophila_CYP4E2 | * | |
| Drosophila_CYP4E3 | Drosophila_CYP4E3 | * | |
| Drosophila_CYP4G1 | Drosophila_CYP4G1 | * | |
| Drosophila_CYP4G15 | Drosophila_CYP4G15 | * | |
| Drosophila_CYP4P1 | Drosophila_CYP4P1 | * | |
| Drosophila_CYP4P2 | Drosophila_CYP4P2 | * | |
| Drosophila_CYP4P3 | Drosophila_CYP4P3 | * | |
| Drosophila_CYP4S3 | Drosophila_CYP4S4 | * | |
| **Fungi** | | | |
| Choanoflagellates_gi_167526307_ref_XP_001747 | Choanoflagellates_XP_001747487 | XP_001747487 | |
| Fungi_CYP_gi_238490512_ref_XP_0023 | Fungi_CYP_XP_002376493 | XP_002376493 | |
| Fungi_CYP_maybe_gi_242779870_ | Fungi_CYP_XP_002479477 | XP_002479477 | |
| Fungi_CYP4 | Fungi_CYP4_XP_001820767 | XP_001820767 | |

* Sequences retrieved from the P450 Homepage

**Table A2.2. Evolutionary functional divergence (Type I) of the CYP4 subfamilies based on taxonomic class.** Assessment of the site-specific evolutionary rates of 13 CYP4 subfamilies group comparisons (with more than 4 gene representatives) via DIVERGE[a]. 13 groups were identified for the type I functional divergence analysis; these groups are specifc for subfamily and taxonimic class.

| Subfamily groups | 4A mam[b] | 4X mam[b] | 4B mam[b] | 4T amp[c] | 4T act[d] | 4F mam[b] | 4F act[d] | 4F22 mam[b] | 4V mam[b] | 4V act[d] | 4* ech[e] | 4D ins[f] | 4*[g] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4A mam** | ---------- | 0.55 ± 0.111 | 0.299 ± 0.081 | 0.491 ± 0.121 | 0.388 ± 0.096 | 0.461 ± 0.070 | 0.297 ± 0.086 | 0.468 ± 0.236 | 0.762 ± 0.109 | 0.685 ± 0.097 | 0.688 ± 0.112 | 0.541 ± 0.098 | 0.434 ± 0.217 |
| **4X mam** | 24.432 | ---------- | 0.65 ± 0.118 | 0.925 ± 0.166 | 0.582 ± 0.150 | 0.541 ± 0.080 | 0.674 ± 0.137 | 0.001 ± 0.022 | 0.966 ± 0.155 | 0.914 ± 0.129 | 0.625 ± 0.156 | 0.638 ± 0.122 | 0.496 ± 0.342 |
| **4B mam** | 13.629 | 30.358 | ---------- | 0.526 ± 0.156 | 0.233 ± 0.088 | 0.527 ± 0.077 | 0.442 ± 0.124 | 0.28 ± 0.217 | 0.639 ± 0.122 | 0.578 ± 0.101 | 0.388 ± 0.119 | 0.546 ± 0.104 | 0.32 ± 0.213 |
| **4T amp** | 16.588 | 30.940 | 11.381 | ---------- | 0.397 ± 0.136 | 0.625 ± 0.102 | 0.322 ± 0.16 | 0.469 ± 0.222 | 0.498 ± 0.122 | 0.596 ± 0.139 | 0.495 ± 0.154 | 0.407 ± 0.117 | 0.753 ± 0.361 |
| **4T act** | 16.352 | 15.097 | 6.927 | 8.457 | ---------- | 0.54 ± 0.080 | 0.454 ± 0.118 | 0.247 ± 0.229 | 0.5 ± 0.110 | 0.608 ± 0.104 | 0.515 ± 0.125 | 0.64 ± 0.139 | 0.601 ± 0.295 |
| **4F mam** | 42.801 | 45.680 | 46.674 | 37.674 | 45.328 | ---------- | 0.279 ± 0.072 | 0.337 ± 0.179 | 0.614 ± 0.088 | 0.494 ± 0.071 | 0.422 ± 0.074 | 0.394 ± 0.066 | ND |
| **4F act** | 11.860 | 24.301 | 12.756 | 4.064 | 14.905 | 14.905 | ---------- | 0.101 ± 0.338 | 0.396 ± 0.122 | 0.66 ± 0.111 | 0.476 ± 0.131 | 0.372 ± 0.091 | 0.195 ± 0.167 |
| **4F22 mam** | 3.939 | 0.000 | 1.672 | 4.465 | 1.160 | 3.531 | 0.089 | ---------- | 0.304 ± 0.305 | 0.317 ± 0.271 | 0.519 ± 0.306 | 0.001 ± 0.022 | 0.001 ± 0.022 |
| **4V mam** | 49.037 | 38.731 | 27.372 | 16.629 | 20.762 | 49.182 | 10.575 | 0.994 | ---------- | 0.016 ± 0.089 | 0.314 ± 0.118 | 0.314 ± 0.119 | 0.485 ± 0.268 |
| **4V act** | 50.309 | 50.586 | 32.794 | 18.464 | 34.082 | 49.052 | 35.059 | 1.366 | 0.033 | ---------- | 0.334 ± 0.103 | 0.187 ± 0.098 | 0.33 ± 0.238 |
| **4* ech** | 37.531 | 16.111 | 10.543 | 10.355 | 17.029 | 32.826 | 13.105 | 2.871 | 7.150 | 10.448 | ---------- | 0.313 ± 0.102 | 0.001 ± 0.022 |
| **4D ins** | 30.436 | 27.430 | 27.598 | 12.012 | 21.112 | 35.352 | 16.641 | 0.000 | 6.930 | 3.664 | 9.341 | ---------- | ND |
| **4** | 3.999 | 2.107 | 2.261 | 4.338 | 4.137 | ND | 1.372 | 0.000 | 3.270 | 1.931 | 0.000 | ND | ---------- |

[a] - Measured theta (θ ), coefficient of evolutionary functional divergence, and their respective standard error are shown above the diagonal. Functional divergence values are represented by $1 > \theta > 0.5$. The likelihood ratio test (LRT) values are shown bellow the diagonal. The LRT tested the null hypothesis of $\theta = 0$. Highlighted LRT values significantly reject the null hypothesis ($p < 0.05$).

[b] - Mammalia

[c] - Amphibia

[d] - Actinopterygii

[e] - Echinoidia

[f] - Insecta

[g] - Fungi & Choanoflagellates

* - these CYP4 sequences have not been assigned to a specific subfamilyputative CYP4 sequences

ND - No detectable data

**Table A2.3. Sites of radical change in the CYP 4F22/ 4F*/4F fish taxonomic groups.** Sites of radical change were detected in pairwise type II functional analyses of CYP4F* (mammals excluding CYP4F22 mammalian clade), CYP4F fish, and CYP4F22 groups. The residue positions are provided based on a reference sequence, human CYP4F22. The amino acid changes between subfamilies and the type of change are provided.

| Residue position (Human CYP4F22) | 4F22/4F* | 4F22/4F fish | Amino Acid Changes | Helix | Property Changes[1] |
|---|---|---|---|---|---|
| 37[2] [188][3] | 1.258 | | R->W | before A | +/Hydrophobic |
| 108 [330] | | 1.411 | V->R | between A and B | Hydrophobic/+ |
| 126 [359] | | 1.411 | A->T | between B and B' | Hydrophobic/Hydrophilic |
| 212 [462] | | 1.411 | Q->L | E | Hydrophilic/Hydrophobic |
| 341 [737] | 3.442 | 1.411 | S->A | I | Hydrophilic/Hydrophobic |
| 354 [750] | 3.442 | | Y->H | J' | Hydrophobic/+ |
| 363 [759] | 1.258 | | E->Q | J' | -/Hydrophilic |
| 388 [787] | 3.442 | | T->L | K | Hydrophilic/Hydrophobic |
| 398 [797] | 1.262 | 1.411 | Q->L | K | Hydrophilic/Hydrophobic |
| 399 [798] | 1.168 | | F,H[4]->H | K | Hydrophobic,+/+ |
| 409 [808] | | 1.411 | C->Y | between K and K' | Hydrophilic/Hydrophilic |
| 417 [816] | | 1.411 | D->G | between K and K' | -/Hydrophilic |
| 445 [898] | 3.442 | | K->E | between K and K' | +/- |
| 463 [919] | | 1.411 | L->H | Pass K" | Hydrophobic/+ |
| 499 [982] | 1.288 | | S->L/F[5] | Pass L | Hydrophilic/Hydrophobic |
| 509 [1046] | | 1.411 | K->L | Pass L | +/Hydrophobic |
| 511 [1048] | | 1.411 | E->Q | Pass L | -/Hydrophilic |

[1]-Property changes are given for the first/second taxonomic group in the analyses, + is positively charged, - is negatively charged

[2]-Residue position based on a reference sequence, Human CYP4F22

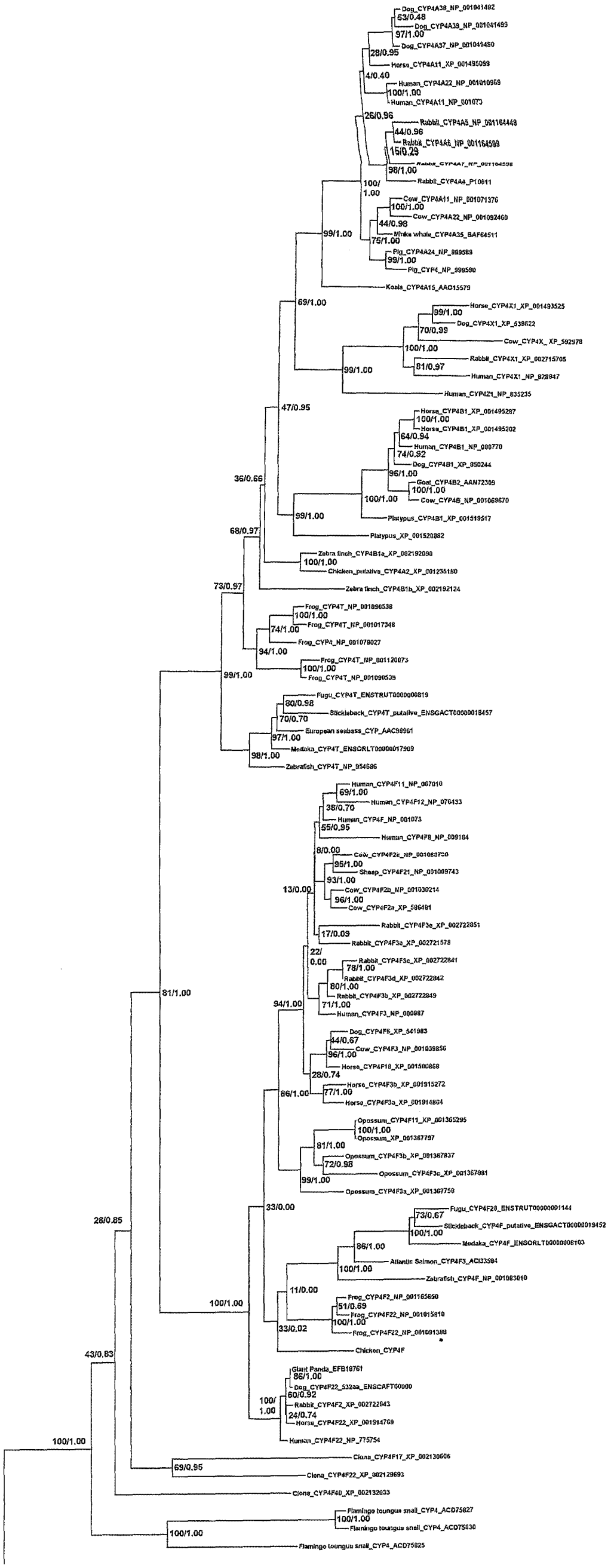
[3]-Indicates the position in the full CYP4 alignment

[4]-Human specific amino acid

[5]-Opossum specific amino acid

## Appendix 2 Figures

**Figure A2.1. Detailed cytochrome P450 family 4 phylogeny.** The cytochrome P450 (CYP) 4 family phylogenetic reconstruction showing the evolutionary relationships between the CYP4 subfamilies from invertebrates and vertebrate species. The phylogeny was determined by maximum likelihood and a WAG+I+G+F (Whelan and Goldman 2001) model (see materials and methods for details of phylogenetic methods). All nodes show phylogenetic support as bootstrapping values / posterior probabilities.


Dog_CYP4A38_NP_001041492
53/0.48
Dog_CYP4A39_NP_001041499
97/1.00
Dog_CYP4A37_NP_001041490
28/0.95
Horse_CYP4A11_XP_001495099
4/0.40
Human_CYP4A22_NP_001010969
100/1.00
Human_CYP4A11_NP_001073
26/0.96
Rabbit_CYP4A5_NP_001164448
44/0.96
Rabbit_CYP4A6_NP_001164599
15/0.29
96/1.00
Rabbit_CYP4A4_P10611
100/
1.00
Cow_CYP4A11_NP_001071376
100/1.00
Cow_CYP4A22_NP_001092460
44/0.98
Minke whale_CYP4A35_BAF64511
75/1.00
Pig_CYP4A24_NP_999589
99/1.00
Pig_CYP4_NP_999590
99/1.00
Koala_CYP4A15_AAO15579
69/1.00
Horse_CYP4X1_XP_001493525
99/1.00
Dog_CYP4X1_XP_539622
70/0.99
Cow_CYP4X_ XP_592978
100/1.00
Rabbit_CYP4X1_XP_002715705
81/0.97
Human_CYP4X1_NP_828847
99/1.00
Human_CYP4Z1_NP_835235
47/0.95
Horse_CYP4B1_XP_001495287
100/1.00
Horse_CYP4B1_XP_001495202
64/0.94
Human_CYP4B1_NP_000770
74/0.92
Dog_CYP4B1_XP_850244
96/1.00
Goat_CYP4B2_AAN72309
100/1.00
Cow_CYP4B_NP_001069670
36/0.66
100/1.00
Platypus_CYP4B1_XP_001519517
99/1.00
Platypus_XP_001520882
68/0.97
Zebra finch_CYP4B1a_XP_002192090
100/1.00
Chicken_putative_CYP4A2_XP_001235180
73/0.97
Zebra finch_CYP4B1b_XP_002192124
Frog_CYP4T_NP_001090538
100/1.00
Frog_CYP4T_NP_001017348
74/1.00
Frog_CYP4_NP_001079027
94/1.00
Frog_CYP4T_NP_001120073
100/1.00
Frog_CYP4T_NP_001090539
99/1.00
Fugu_CYP4T_ENSTRUT00000000819
80/0.98
Stickleback_CYP4T_putative_ENSGACT00000018457
70/0.70
European seabass_CYP_AAC98961
97/1.00
Medaka_CYP4T_ENSORLT00000017909
98/1.00
Zebrafish_CYP4T_NP_954686
Human_CYP4F11_NP_067010
69/1.00
Human_CYP4F12_NP_076433
38/0.70
Human_CYP4F_NP_001073
55/0.95
Human_CYP4F8_NP_009184
8/0.00
Cow_CYP4F2c_NP_001068750
95/1.00
Sheep_CYP4F21_NP_001009743
93/1.00
13/0.00
Cow_CYP4F2b_NP_001030214
96/1.00
Cow_CYP4F2a_XP_586491
Rabbit_CYP4F3c_XP_002722851
17/0.09
Rabbit_CYP4F3a_XP_002721578
22/
0.00
Rabbit_CYP4F3c_XP_002722841
78/1.00
Rabbit_CYP4F3d_XP_002722842
80/1.00
Rabbit_CYP4F3b_XP_002722849
71/1.00
Human_CYP4F3_NP_000887
94/1.00
Dog_CYP4F5_XP_541983
44/0.67
Cow_CYP4F3_NP_001039856
96/1.00
Horse_CYP4F18_XP_001500868
28/0.74
Horse_CYP4F3b_XP_001915272
77/1.00
86/1.00
Horse_CYP4F3a_XP_001914864
Opossum_CYP4F11_XP_001365295
100/1.00
Opossum_XP_001367797
81/1.00
Opossum_CYP4F3b_XP_001367837
72/0.98
Opossum_CYP4F3c_XP_001367881
99/1.00
Opossum_CYP4F3a_XP_001367758
33/0.00
Fugu_CYP4F28_ENSTRUT00000001144
73/0.67
Stickleback_CYP4F_putative_ENSGACT00000019452
100/1.00
Medaka_CYP4F_ENSORLT00000006103
28/0.85
86/1.00
Atlantic Salmon_CYP4F3_ACI33584
100/1.00
Zebrafish_CYP4F_NP_001083010
11/0.00
Frog_CYP4F2_NP_001165650
100/1.00
51/0.69
Frog_CYP4F22_NP_001015810
33/0.02
Frog_CYP4F22_NP_001091388
Chicken_CYP4F
43/0.83
Giant Panda_EFB18761
86/1.00
Dog_CYP4F22_532aa_ENSCAFT00000
60/0.92
100/
1.00
Rabbit_CYP4F2_XP_002722843
24/0.74
Horse_CYP4F22_XP_001914769
Human_CYP4F22_NP_775754
100/1.00
Ciona_CYP4F17_XP_002130506
69/0.95
Ciona_CYP4F22_XP_002129693
Ciona_CYP4F40_XP_002132033
Flamingo tounge snail_CYP4_ACD75827
100/1.00
Flamingo tounge snail_CYP4_ACD75830
100/1.00
Flamingo tounge snail_CYP4_ACD75825

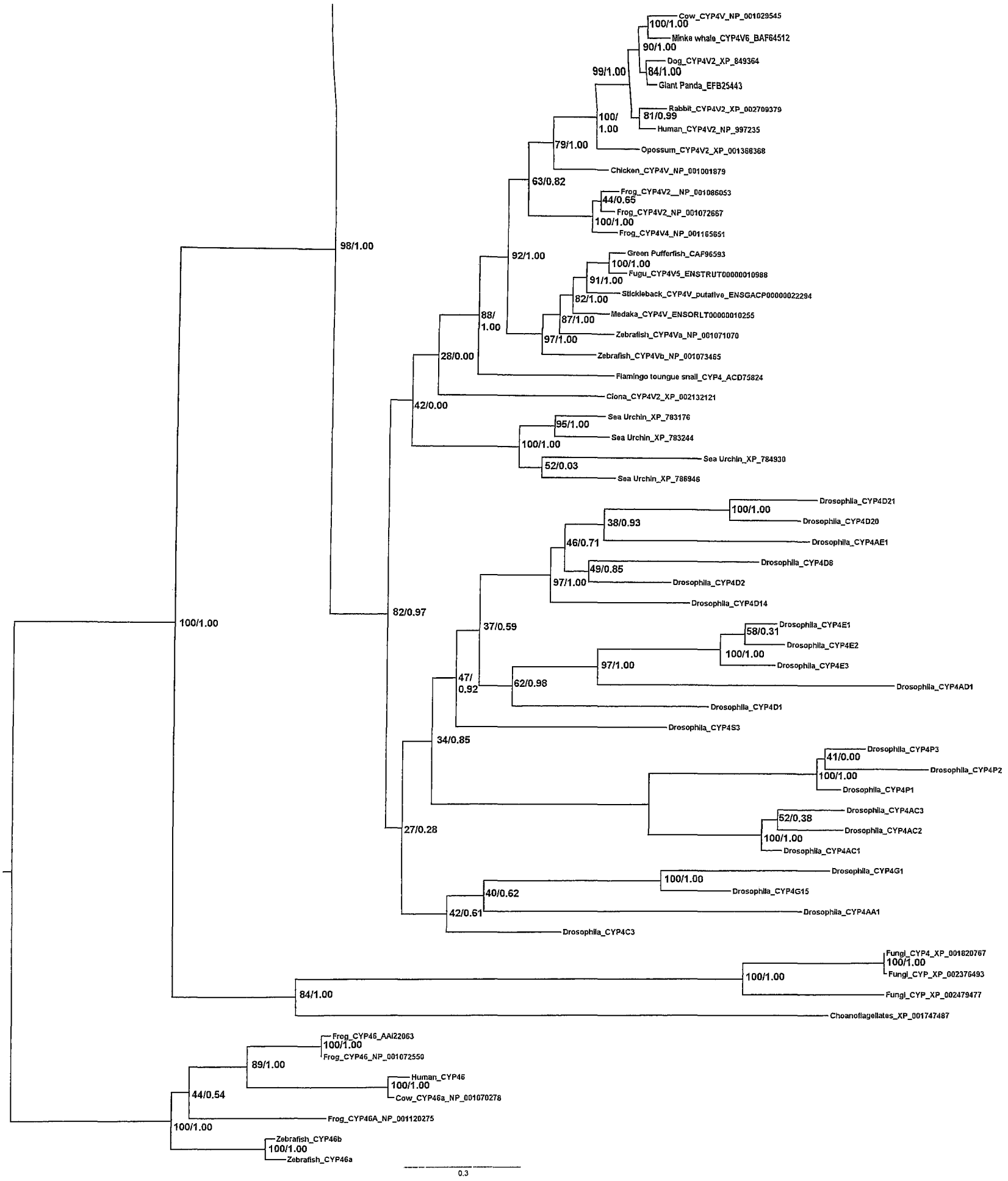
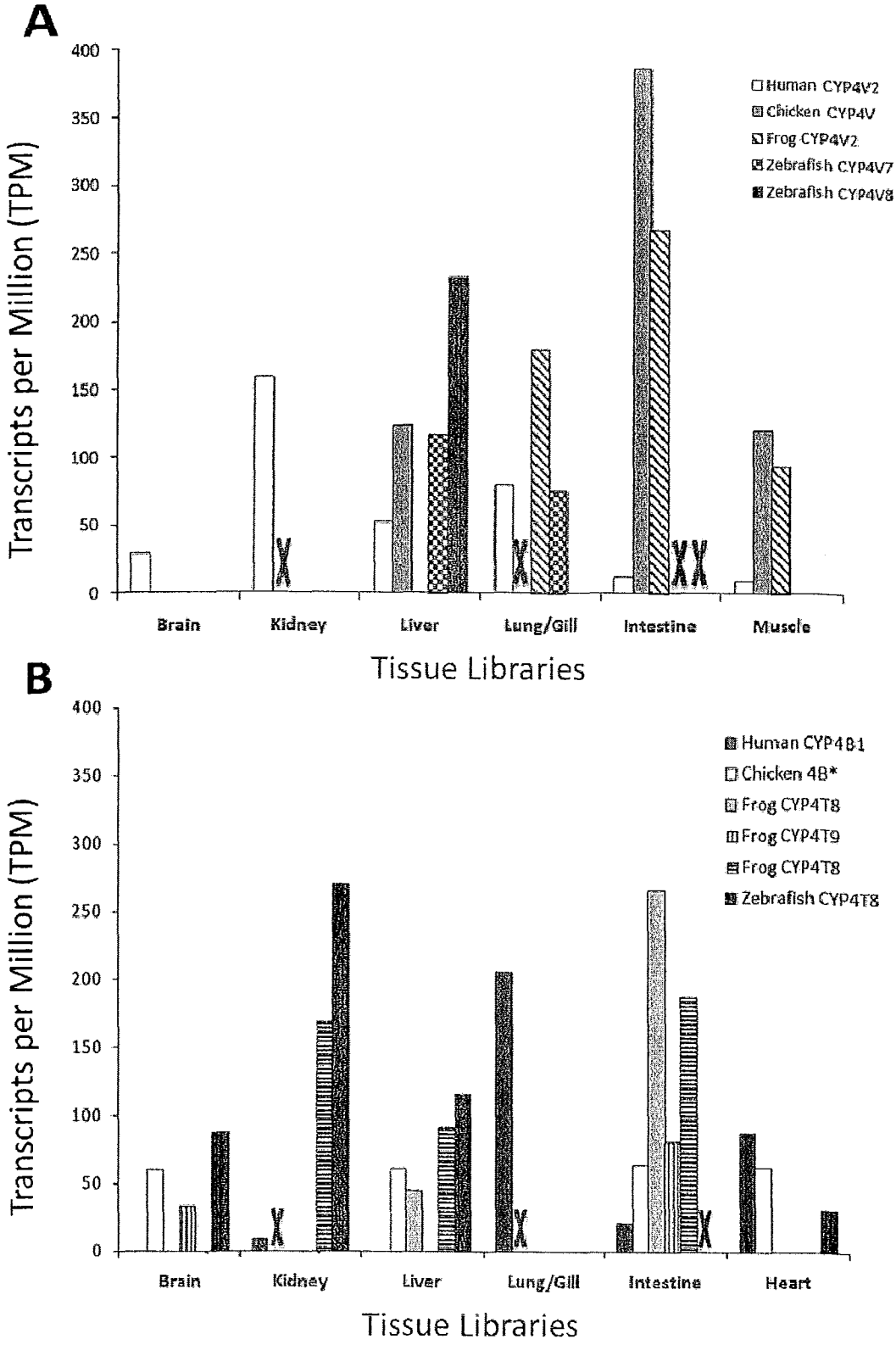

Cow_CYP4V_NP_001029545
100/1.00
Minke whale_CYP4V6_BAF64512
90/1.00
Dog_CYP4V2_XP_849364
84/1.00
Giant Panda_EFB25443
99/1.00
Rabbit_CYP4V2_XP_002709379
81/0.99
Human_CYP4V2_NP_997235
100/
1.00
Opossum_CYP4V2_XP_001368368
79/1.00
Chicken_CYP4V_NP_001001879
63/0.82
Frog_CYP4V2__NP_001086053
44/0.65
Frog_CYP4V2_NP_001072667
100/1.00
Frog_CYP4V4_NP_001165651
98/1.00
Green Pufferfish_CAF96593
92/1.00
100/1.00
Fugu_CYP4V5_ENSTRUT00000010988
91/1.00
Stickleback_CYP4V_putative_ENSGACP00000022294
82/1.00
Medaka_CYP4V_ENSORLT00000010255
88/
1.00
87/1.00
Zebrafish_CYP4Va_NP_001071070
97/1.00
Zebrafish_CYP4Vb_NP_001073465
28/0.00
Flamingo tounge snail_CYP4_ACD75824
Ciona_CYP4V2_XP_002132121
42/0.00
Sea Urchin_XP_783176
95/1.00
Sea Urchin_XP_783244
100/1.00
Sea Urchin_XP_784930
52/0.03
Sea Urchin_XP_786946
Drosophila_CYP4D21
100/1.00
38/0.93
Drosophila_CYP4D20
Drosophila_CYP4AE1
46/0.71
Drosophila_CYP4D8
49/0.85
97/1.00
Drosophila_CYP4D2
Drosophila_CYP4D14
82/0.97
37/0.59
Drosophila_CYP4E1
58/0.31
Drosophila_CYP4E2
100/1.00
Drosophila_CYP4E3
97/1.00
47/
0.92
62/0.98
Drosophila_CYP4AD1
Drosophila_CYP4D1
Drosophila_CYP4S3
34/0.85
Drosophila_CYP4P3
41/0.00
Drosophila_CYP4P2
100/1.00
Drosophila_CYP4P1
Drosophila_CYP4AC3
52/0.38
Drosophila_CYP4AC2
100/1.00
Drosophila_CYP4AC1
27/0.28
Drosophila_CYP4G1
100/1.00
Drosophila_CYP4G15
40/0.62
42/0.61
Drosophila_CYP4AA1
Drosophila_CYP4C3
Fungi_CYP4_XP_001820767
100/1.00
100/1.00
Fungi_CYP_XP_002376493
Fungi_CYP_XP_002479477
84/1.00
Choanoflagellates_XP_001747487
100/1.00
Frog_CYP46_AAI22063
100/1.00
Frog_CYP46_NP_001072550
89/1.00
Human_CYP46
100/1.00
Cow_CYP46a_NP_001070278
44/0.54
Frog_CYP46A_NP_001120275
100/1.00
Zebrafish_CYP46b
100/1.00
Zebrafish_CYP46a
0.3

**Figure A2.2. Digital gene expression of the vertebrate CYP 4B, 4T, and 4V subfamilies.** Gene expression levels of CYP4 genes are provided, in transcripts per million (TPM), from tissue libraries containing >10,000 ESTs for vertebrate species. Libraries that were not available or had less than 10,000 ESTs for the species of interest are denoted with an X. The digital gene expression patterns are shown in (A) for human, chicken, frog and zebrafish genes in the CYP4V subfamily and (B) for zebrafish and frog genes in CYP4T and chicken and human genes in CYP4B subfamilies. The chicken 4B* is the CYP4A/4B sequence from Genbank that clustered with the CYP4B subfamily, thus it is noted with an astrix.


A
Transcripts per Million (TPM)
400
350
300
250
200
150
100
50
0
Human CYP4V2
Chicken CYP4V
Frog CYP4V2
Zebrafish CYP4V7
Zebrafish CYP4V8
Brain
Kidney
Liver
Lung/Gill
Intestine
Muscle
Tissue Libraries
B
Transcripts per Million (TPM)
400
350
300
250
200
150
100
50
0
Human CYP4B1
Chicken 4B*
Frog CYP4T8
Frog CYP4T9
Frog CYP4T8
Zebrafish CYP4T8
Brain
Kidney
Liver
Lung/Gill
Intestine
Heart
Tissue Libraries

# APPENDIX 3:

# MICROARRAY CHIP CUSTOM CYP AND NUCLEAR RECEPTOR PROBES FOR THE ZEBRAFISH GENOME

## Custom Probe design for a Zebrafish (Zv8) Microarray Chip

The available zebrafish microarray chip from Agilent technologies presents a good representation of the zebrafish genome, yet the CYP genes are not fully represented. Since the microarray would be used to study differential gene expression with pharmaceutical exposures, the CYP and nuclear receptor genes coverage was assessed and custom probes designed to provide full coverage. The custom probes were designed based on the zebrafish genome, ensemble 56 from the assembly Zv8. All CYP and nuclear receptor genes were targeted during the custom probe design process. Zebrafish CYP genes were acquired from the P450 homepage and as well as the Ensembl data base. The sequences from the P450 homepage were BLAST searched against the Ensembl database in order to generate a complete list of the genes. A total of 146 CYP probes were found on the Agilent chip; 16 probes were not unique. For the genes that lacked a unique probe, custom probes were designed in e-array website (https://earray.chem.agilent.com/earray/). Three custom probes were made for nuclear receptors not seen represented within the existing probe set. The design of the probes favored the 3' end of the transcript with a 60-mer length. The custom probes including their sequence and annotation are provided in table A3.1.

**Table A3.1:** Custom WilsonLab CYP and nuclear receptor microarray probes for zebrafish genome. All probe names, gene symbols and probe sequences are represented in the table.

| Probe_name | Gene_symbol | Probe Sequence | Annotation |
|---|---|---|---|
| WilsonLab_11 | cyp2k-like | TCGTCTGTTTGTGGCAAATCAAAAGCGTGTTGTGGACAATGTCCAAGAATCCTTTAAACA | cytochrome P450, family 2, subfamily K, like |
| WilsonLab_13 | cyp2p12-putative | TACTTCAGAAGATACTCAGATAGAAAAATACTCTATTCCAAAGGGCACAATGGTGACCAG | cytochrome P450, family 2, subfamily P, polypeptide 12, putative |
| WilsonLab_17 | cyp2x8 | TTTGATTACAACAGTGAAACCCTTCAGTGCTACATTCAGCTCATTACCGAGATTTCAAAG | cytochrome P450, family 2, subfamily X, polypeptide 8 |
| WilsonLab_2 | cyp2aa7 | ACAAACCTGGCTGCTATTCTAAGCGATAAGGAGCACTGGAAGCATCCTGACACATTTAAC | cytochrome P450, family 2, subfamily AA, polypeptide 7 |
| WilsonLab_22 | cyp2y3-putative | AGAAGTTCACCTTCAGTAGTCCTAATGGTCCAGATGGAATTGACCTCAGTCCTGAACTCA | cytochrome P450, family 2, subfamily Y, polypeptide 3, putative |
| WilsonLab_25 | esrrgb | ATGTTTGCACATCTGAATAATCTCCCCTGTCTGCATGTGTCTGCTGTAGTGACGAAGATC | estrogen-related receptor gamma b |
| WilsonLab_28 | pparab | AAACCTCGACTTAAACGATCAGGTAACGTTACTGAAGTACGGTGTCCATGAGGCTCTGTT | peroxisome proliferator activated receptor alpha b |
| WilsonLab_31 | cyp2p8 | TATAAAATGGGGGGCACACACTGTCCTAAGCCATTTAAACTGTGTGCAGTTCCACGCTAA | cytochrome P450, family 2, subfamily P, polypeptide 8 |
| WilsonLab_34 | cyp2x11 | ATATATATGGAGGGACTCAGTCATTGAAACCTTACCCTATGATTGTAGAGCTGCGGACGC | cytochrome P450, family 2, subfamily X, polypeptide 11 |
| WilsonLab_37 | cyp2x10 | TATCAGAGTTCTGCACGGTCTCCTCTCCACCTTTCACAACAAATAAAAGCTATTGAATTA | cytochrome P450, family 2, subfamily X, polypeptide 10 |
| WilsonLab_4 | cyp2k18 | TCAAAGGAATTGTTGGAATCACGTTGAACCCATCTCCACACAAGCTGTGTGCAATCAGAC | cytochrome P450, family 2, subfamily K, polypeptide 18 |
| WilsonLab_46 | cyp2x6 | CATCACAGTTTTGTGAAATAAATCTTTTTGCACATCTCAGATGTTCAGTTTGAAATGCGC | cytochrome P450, family 2, subfamily X, polypeptide 6 |
| WilsonLab_49 | cyp7a | CTGGGTGTACTGCAGCCCACCTATGATGTTGACTTTCGTTACAGACTCAAATCTCTCTAA | cytochrome P450, family 7, subfamily A |
| WilsonLab_52 | cyp11a1 | CTGGATCTGTTATTGTGTGTGAGTTGGGCACTATAATATGAGACCATGATCCTTTTAAAT | cytochrome P450, subfamily XIA, polypeptide 1 |
| WilsonLab_55 | cyp11b2 | AGCACTAAATACACACTGATCCTGCAGCCCGAGTGTCCACCGAGAATCACCTTCAGCACA | cytochrome P450, family 11, subfamily B, polypeptide 2 |
| WilsonLab_58 | cyp20 | GGCCTGCCCTGAATTAAGGTATGAGGAGAGAAATACAATCAAAACAGTTTTATTAAATCT | cytochrome P450, family 20 |
| WilsonLab_61 | cyp26b1 | GTGCCATTTGACTACATGTTCTGTTCTTTTGTCGATTGTTGCGTTTGTCTGTACATATAT | cytochrome P450, family 26, subfamily b, polypeptide 1 |
| WilsonLab_64 | cyp27b1 | ACATCGAAGTCGGAGGATACGTCATTCCTAAAAACACTCTGATCACGCTCTGTCATTACG | cytochrome P450, family 27, subfamily B, polypeptide 1 |
| WilsonLab_70 | rargl | TAACAGAGCTGGTATCTAAGATGAGGGAAATGATGGACAAGACGGAGCTGGGCTGTTTAA | retinoic acid receptor gamma, like |
| WilsonLab_9 | cyp2k20-putative | GTGCCTTTGAGTCTTCCACATAGAACTACCAGTGACATTACCTTCAATGGATACTTCATC | cytochrome P450, family 2, subfamily K, polypeptide 20, putative |