Isolation and phylogeny of novel cytochrome P450 genes from tunicates (*Ciona* spp.): A CYP3 line in early deuterostomes?

TIM VERSLYCKE[1*], JARED V. GOLDSTONE[1] AND JOHN J. STEGEMAN[1]

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

* Corresponding author:

Tim Verslycke

Biology Department MS#32,

Woods Hole Oceanographic Institution,

Woods Hole, MA 02543, USA.

tel : +1 508 289 3729

fax : +1 508 547 2134

e-mail : tim@whoi.edu

Abstract

Cytochromes P450 (CYPs) form a gene superfamily involved in the biotransformation of numerous endogenous and exogenous natural and synthetic compounds. In humans, CYP3A4 is regarded as one of the most important CYPs due to its abundance in liver and its capacity to metabolize more than 50% of all clinically used drugs. It has been suggested that all CYP3s arose from a common ancestral gene lineage that diverged between 800 and 1100 million years ago, before the deuterostome-protostome split. While CYP3s are well known in mammals and have been described in lower vertebrates, they have not been reported in non-vertebrate deuterostomes. Members of the genus *Ciona* belong to the tunicates, whose lineage is thought to be the most basal among the chordates, and from which the vertebrate line diverged. Here we describe the cloning, exon-intron structure, phylogeny, and estimated expression of four novel genes from *Ciona intestinalis*. We also describe the gene structure and phylogeny of homologous genes in *Ciona savignyi*. Comparing these genes with other members of the CYP clan 3, show that the *Ciona* sequences bear remarkable similarity to vertebrate CYP3A genes, and may be an early deuterostome CYP3 line.

Keywords: *Ciona intestinalis*, *Ciona savignyi*, Cytochrome P450, CYP3, phylogeny, nomenclature.

1. Introduction

The cytochrome P450 (CYP) gene superfamily is significant in a wide variety of disciplines, ranging from medical genetics to inorganic chemistry (Guengerich, 1991), due to CYP roles in oxidative transformation of exogenous and endogenous organic compounds. The natural history of CYP dates back perhaps 3 billion years, and they occur variously in archaea, eubacteria and eukaryota (Nelson et al., 1996; Stegeman and Livingstone, 1998). Currently, CYP sequences are classified into families and subfamilies based on percent identity (Nelson et al., 1996). However, relationships among many CYPs within vertebrate taxa, and more so between vertebrates and invertebrates, are obscured because of evolutionary distance, independent duplications, gene conversion and gene loss. A higher order clustering of gene families into “clans” (Nelson, 1998) has also been proposed, to indicate probable common evolutionary origin of those gene families. Yet, the number of novel CYPs identified increases greatly as genome sequencing addresses an expanding number of taxa, which exacerbates the challenges of discerning CYP relationships and nomenclature. CYPs in species representing phylogenetic lines that diverged at key intervals should help elucidate the origin of CYP lineages and functions, and shed light on the significance of CYP diversity in animals.

The origin of CYP families in vertebrates may be clarified by studies on early diverging deuterostomes, such as the non-vertebrate chordates belonging to the subphylum *Tunicata* (cf. *Urochordata*; Zeng and Swalla, 2005). The tunicate lineage is believed to be the most basal among the chordates, diverging prior to the cephalochordates and the vertebrates (Schaeffer, 1987), although see Graham (2004) and Delsuc et al. (2006). Because of this evolutionary position, their simple and short embryonic development, the large-scale gene duplications that occurred in the vertebrate lineage after divergence from urochordates and cephalochordates, and their small genome size (an estimated 160 million base pairs), the U.S. Department of Energy Joint Genome Institute (JGI) selected the ascidian *Ciona intestinalis* for a genome sequencing project (Dehal et al., 2002). The genome of the congeneric species *Ciona savignyi* has been sequenced and the genome assembly released in 2003 (http://www.broad.mit.edu/annotation/ciona/index). Our efforts to describe CYPs in these *Ciona* species have focused on possible members of the CYP1 and CYP3 families.

The CYP3 family, specifically the CYP3A subfamily, has been studied with great intensity because of its substrate diversity, and importance in drug development and discovery. The CYP3As catalyze the metabolism of 40 to 60% of all clinically used drugs in humans (Guengerich, 1999), and also metabolize endogenous hormones, bile acids, fungal and plant products, and environmental pollutants (reviewed by Maurel, 1996; Thummel and Wilkinson, 1998; Guengerich, 1999). The diversity of genes in some vertebrate CYP families, e.g., the CYP2s, is extreme. Recent molecular phylogenetic studies explored relationships among CYP3 genes between vertebrates (McArthur et al., 2003; Williams et al., 2004), showing that the CYP3s are much more conserved than the CYP2s, undergoing diversification within taxa, yet with the great majority of genes retaining sufficient identity to be classified in the same subfamily. The biochemistry of CYP3As also is broadly similar from fish to mammals (Celander et al., 1996; Hegelund and Celander, 2003; Bresolin et al., 2005). It has been suggested that the CYP3 ancestral line arose in a bilaterian ancestor between 800 and 1100 million years ago (mya) (Nebert and Gonzalez, 1987; Gonzalez, 1990), but there have been no CYP3s described in non-vertebrate species. Because of the possible conservation of CYP3s in the chordate lineage, we sought possible homologues in the *Tunicata*.

In the present study, we used a CYP3A profile Hidden Markov Model (HMM; Eddy,

1998) to search the *C. intestinalis* predicted protein set for CYP3A-like genes. This HMM search led to the identification and subsequent cloning and sequencing of four novel CYP3A-like genes in *C. intestinalis*. We also investigated homologous CYP genes found in the congeneric ascidian *Ciona savignyi*. The phylogenetic relationships and gene structure comparisons indicate that these *Ciona* CYPs are members of the CYP clan 3 closely resembling the vertebrate CYP3s, leading us to suggest that they represent a tunicate CYP3 line. Finally, an existing *C. intestinalis* cDNA projects database was searched for the presence of the four novel CYP3-like genes indicating variable developmental and tissue expression.

2. Material and Methods

2.1. *Ciona* genome searches

The *C. intestinalis* predicted protein set (http://www.jgi.doe.gov/ciona; Dehal et al., 2002) was searched for CYP3A-like sequences using a Hidden Markov Model constructed from four CYP3A genes using Hmmer 2.3 (Eddy, 1998). These models are significantly more sensitive than single query BLAST searches, as this technique statistically models the gene structure of interest based on previously identified training sequences. The *C. savignyi* genome (http://www.broad.mit.edu/annotation/ciona/) was searched using BLAST for genes similar to the identified *C. intestinalis* CYP3A-like genes, as no global gene prediction set is available for the *C. savignyi* genome. All gene predictions were refined using the HMM protein homolog-based gene prediction software FGENESH+ (Solovyev and Salamov, 1999).

2.2. Primer design and RT-PCR

Gene-specific primers (internal and full-length) were designed for all four genes based on the *C. intestinalis* assembly (v1.95) gene predictions. Primers were designed using MacVector (Accelrys, Inc.). Primers, expected cDNA length, and PCR conditions are depicted in the supplementary material (Supplementary material Table 1). Total RNA was isolated from adult animals using Stat-60 (Tel-Test, Inc., Friendship, TX). Complementary DNA was synthesized from 2 μg of total RNA using random hexamers and the RT-PCR Omniscript cDNA Synthesis Kit (Qiagen). We used the BD Advantage^{TM} 2 PCR enzyme system (BD Biosciences, Clontech) with a Perkin-Elmer GeneAmp 2400 thermocycler.

2.3. Cloning and Sequencing

All PCR products were cloned into pGEM-T Easy (Promega) and sequenced at the Josephine Bay Paul Center Sequencing Facility (Marine Biological Laboratory, Woods Hole, MA). Both strands from multiple clones (at least 6 forward and 6 reverse reads per sequence) were sequenced to ensure accuracy. DNA reads were analyzed, assembled, and translated using Sequencher^{TM} (Gene Codes Corperation). After comparison with the gene predictions from the *C. intestinalis* genome assembly and sequence data from the cDNA projects website, consensus sequences were obtained for all four novel *Ciona* CYP genes.

2.4. Sequence alignments and phylogeny

All sequence alignments were done using ClustalX (Thompson et al., 1997). The final amino acid alignment included 30 taxa with 627 characters, of which 345 were unambiguously aligned, manually masked, and used in further analyses (full alignment is available upon request). Phylogenetic relationships were investigated using Bayesian techniques as implemented in the computer program MrBayes v 3.1.1 (Ronquist and Huelsenbeck, 2003).

MrBayes estimates posterior probabilities of clade support using Metropolis-coupled Monte Carlo Markov Chain method (MC^3). We performed MC^3 estimates with uninformative prior probabilities using the WAG model of amino acid substitution (Whelan and Goldman, 2001) and prior uniform gamma distributions approximated with four categories (WAG+I+Γ), as suggested by analysis of the alignment with ProtTest (v1.2.6; Abascal et al., 2005). Four incrementally heated, randomly seeded Markov chains were run for 10^7 generations and topologies were sampled every 100th generation. Analysis of the MC^3 parameter output using Bayesian Output Analysis (BOA v1.71; http://www.public-health.uiowa.edu/boa/) indicated that this degree of sampling was sufficient to avoid significant sampling autocorrelation. Four independent sets of analyses were performed, starting from random trees. Convergence of all nodes to stationarity was determined using the online version of AWTY (Wilgenbusch et al., 2004) to occur prior to the 10^6 generations discarded as MC^3 burnin. Posterior probabilities of clades were estimated from the sampled topologies after removal of the initial MC^3 burnin. Bayes factors are defined as the ratio of the posterior to the prior odds for the two hypotheses in question (Kass and Raftery, 1995; Huelsenbeck and Immenov, 2002; Suchard et al., 2005). In testing of the monophyly of certain clades within the same tree, the model prior odds are the same and thus the Bayes factor is computed as the ratio of the frequencies of the two hypotheses in the filtered MC^3 run, corrected for the prior number of possible trees. Following Suchard et al. (2005), we have considered the cluster of taxa for which we are testing the hypothesis of monophyly to be rooted within the overall unrooted phylogenetic tree. Topology filtering was performed using PAUP* (v4b10; Swofford, 2003). Maximum likelihood estimates of the topology and branch lengths were obtained using PhyML (Guindon and Gascuel, 2003) with the WAG+I+Γ model approximated with four categories. Estimations of the coefficients of functional divergence and site-specific rate shift analyses were performed with DIVERGE (v1.04; Gu and Vander Velden, 2002).

2.5. Gene conversion

The possible occurrence of gene conversion was tested using the software package RDP2 (Martin et al., 2005). Two simultaneous gene recombination programs (GeneConv and MaxChi) were run using default settings. P-values from individual algorithms were subjected to Bonferroni correction to account for multiple testing and all events with p-value < 0.05 were considered.

2.6. Gene structure comparison

Exon-intron structures were obtained for CYP3A4, CYP6s, and CYP9s (i.e., other clan-3 members) using Ensembl and GenBank, and compared to the exon-intron structure of the *Ciona* CYPs.

2.6. EST database searching and gene location on the chromosomes

The genome sequence of *C. intestinalis* is complemented by extensive expressed sequence tag (EST) analyses of expressed mRNAs within a large-scale EST project in Kyoto, Japan (http://ghost.zool.kyoto-u.ac.jp/indexr1.html; Satou et al., 2002; 2003). This database was searched using the web-based cDNA browser by looking for sequences that correspond to the four *C. intestinalis* CYP3A-like genes. The same web interface was used to determine the different gene locations on the *C. intestinalis* chromosomes (as shown in Figure 1).

3. Results

3.1. Cloning of novel tunicate CYP genes

Profile HMM searching of the *C. intestinalis* predicted protein set resulted in four strong hits (E-value < 10^{-70}). The JGI's gene model numbers (v1.95) for these four predicted genes are ci0100140585 (referred to as CI85 in the rest of the manuscript), ci0100133019 (CI19), ci0100151443 (CI43), and ci0100140050 (CI50). CI50, CI85 and CI43 were found in tandem in the same direction on scaffolds 75 and 41 (found on chromosome 11) of the *C. intestinalis* genome (Figure 1). These sequences code for predicted amino acid sequences of 415 (partial), 399 (partial), and 526 aa (full length), respectively. The fourth gene, CI19, was found on Scaffold 262 (on chromosome 1) as a 1530 bps single-exon sequence coding for a predicted 510 aa fragment.

Specific primers were designed for all four *C. intestinalis* CYPs based on the JGI gene predictions (Supplementary material Table 1). Using an RT-PCR approach, two full-length cDNAs were obtained for CI85 and CI19, coding for 526 and 512 aa, respectively. Partial cDNAs were obtained for CI43 (674 bps fragment) and CI50 (186 bps fragment), demonstrating the expression of all four CYP genes. Full-length consensus sequences were obtained for all four *Ciona* CYPs by comparing the cloned cDNA sequences with the JGI gene model predictions, our own gene predictions from the genomic sequence using Fgenesh+, and the corresponding sequences found through the *C. intestinalis* cDNA/EST database. These sequences (Figure 2) were aligned and used for BLAST searches, and phylogenetic and gene structure analyses.

All four *C. intestinalis* CYPs were used in individual BLAST searches of the *C. savignyi* genome. Four homologous genes were found (referred to as CSV A-D). A BLAST search with the CI19 sequence resulted in the identification of homolog A on contig 3797 of the *C. savignyi* genome (Genbank accession no. **AACT01003797**). Similar to CI19, it consisted of a single exon gene coding for a 512 aa protein. *C. savignyi* homolog B, found on contig 12426 (Genbank accession no. **AACT01012426**), contained 15 exons, and coded for a predicted protein of 515 aa. Homolog C was identified as a predicted fragment of 447 aa containing 13 exons and was found in parts on *C. savignyi* paired scaffolds 126 and 238 (Genbank accession nos. **CH003042** and **CH003161**). Finally, the first 5 exons (a 127 aa fragment) of a fourth homologous gene (homolog D, most similar to CI50) were found on contig 32084 (Genbank accession no. **AACT01032084**). As other exons of homolog D could not be identified in the current *C. savignyi* genome assembly, this gene was not included in the phylogenetic analyses.

3.2. Phylogeny of novel tunicate CYP genes

Figure 3 shows a Bayesian consensus tree containing the novel tunicate CYP sequences and a number of other CYP clan-3 sequences, including a representative selection of CYP3 genes (McArthur et al., 2003; Williams et al., 2004). Four independent replicate analyses provided very similar estimates of posterior probabilities (±0.01). Support for the CYP3 and *Ciona* clades is high, as it is for clustering of the insect-specific CYP6 and CYP9 sequences together. However, posterior probabilities range from 0.60 to 0.81 for the internal branches, and thus the position of the insect CYP6s and CYP9s and vertebrate CYP5A1 with respect to the CYP3 family and the nematode genes is somewhat unclear. Maximum likelihood analysis produced an identical topology and similar branch lengths to the phylogeny obtained using Bayesian methods (see Supplemental Material Figure 1). The clustering of the *Ciona* genes and also the clam *Mercenaria mercenaria* CYP30 sequence with the CYP3 genes is strongly supported (posterior probability of 0.93; Bayes Factor = 15.6). These results point to a decisive conclusion that the *Ciona* genes are phylogenetically similar to the CYP3 genes. The results also are consistent with the original suggestion that the *Mercenaria* CYP30 is related to the vertebrate CYP3s (Brown et al., 1998). The positioning of the *Mercenaria* gene as a sister group to the *Ciona* genes is somewhat suspect, as molluscs would not be expected to cluster with tunicates. However, the correct relative positioning of these genes is likely obscured due to incomplete taxonomic sampling.

The diversity of the novel CYPs in *Ciona* appears to be due to both speciation and to lineage-specific gene duplication. Independent gene duplication events appear to have taken place in each *Ciona* lineage, leading to the observed clustering of the CI85 and CI50 sequences and the respective homologs CSV B and CSV C as separate paralog pairs. In contrast, the single exon CI19 and CSV A genes are orthologues, apparently representing a processed RNA integrated into the genome of an ancestor prior to the divergence of the two *Ciona* species. Unfortunately, due to the current state of the *C. savignyi* genome we were not able to include the fragmentary homolog D in the phylogeny.

Phylogenies resulting in taxon-specific paralog pairs could reflect gene conversion, as seen recently in CYP1A genes (Goldstone and Stegeman, in press). We observed no evidence for gene recombination between *C. savignyi* genes, and only limited evidence for recombination between *C. intestinalis* genes CI43, CI85, and CI50. A stretch of absolute base pair identity between the three genes exists between alignment positions 991-1059, and regions of pairwise identity exist throughout the alignment of these three genes (data not shown). The programs GeneConv and MaxChi both indicate that the 991-1059 region may represent a gene conversion event. However, the overall high identity (72-85%) between the three tandemly duplicated genes (see Table 1 for an amino acid identity; and Supplementary material Table 2 for basepair identity), make it difficult to distinguish between functional conservation and gene conversion.

3.3. Gene structure and similarity within CYP clan 3

Table 1 shows a matrix of overall identities between the four *C. intestinalis* and three *C. savignyi* consensus sequences with selected CYP3 family genes and genes in other families in CYP clan 3. Using current criteria for classification, the *Ciona* sequences all fall within the same subfamily, sharing 52-89% amino acid identity (for the masked alignment). The single exon genes (CI19 and CSV A) are more identical to each other than they are to the other *Ciona* genes, reflected also in the position of these two genes relative to the other five *Ciona* genes in the phylogenetic tree (Figure 3). The shared identities of the masked *Ciona* genes with the CYP3 family genes range from 34-44%, with the highest identities to chicken CYP3A37. Interestingly, the masked clam CYP30 sequence shares a higher percent identity with human CYP3A4 and rat CYP3A9 genes (47 and 45%, respectively) than with the *Ciona* genes (34-39%).

The mammalian CYP3As are characterized by a high degree of structural similarity, with well-conserved exon-intron structures (Hashimoto et al., 1993; Nelson et al., 1996). In this perspective, the exon-intron structure of the *Ciona* CYPs was compared to that of human CYP3A4, and to members of other gene families in CYP clan 3 for which the exon-intron organization is known, i.e., CYP5A1, CYP6s, and CYP9s. As shown in Figure 4, the *Ciona* multiple-exon genes for which we have or can predict full-length sequence, all contain 15 exons with a combined length varying from 1530-1581 bp. Cloning and sequencing confirmed that CI85 and CI43 code for 526 aa and 509 aa full-length CYP proteins, respectively. For CI43, this is 110 aa longer than predicted by the JGI gene model. CI50, as predicted by the JGI gene model, contains 12 exons and runs off the end of Scaffold 75. Our consensus CI50 full-length sequence contains 15 exons with a total length of 1530 bp (coding for a 509 aa full-length CYP protein). CI19 is a single exon gene that codes for a full-length CYP protein that is 512 amino acids in length.

The intergenic regions between genes CI50 and CI85, and between CI85 and CI43 are 550 bp and 772 bp, respectively. Gene clusters have been previously described in the CYP3s and in CYP6s and CYP9s (Finta and Zaphiropoulos, 2000; Ranson et al., 2002). Genes in these clusters are separated by intergenic regions ranging from a few tens, to several thousand bps.

Figure 4 also shows predicted exon-intron boundaries in *C. savignyi* homologs identified in BLAST searches using the *C. intestinalis* sequences. The gene structures are essentially identical to those in *C. intestinalis,* except for minor differences in *Ciona* exons 9, 10 and 15.

The *Ciona* exon lengths are remarkably similar to those of human CYP3A4 (Genbank accession no. **J04449**). A recent phylogenetic study of 42 CYP3A subfamily members found that they all contained 13 exons, with exons 1 and 10 always 71 bp and 161 bp in length, respectively (Williams et al., 2004), corresponding to exons 1 and 11 in the *Ciona* CYP genes in our study. Furthermore, exon 11 in human CYP3A4 is 227 bp in length, which equals the combined length of *Ciona* CYP exons 12 (89 bp) and 13 (138 bp). Exon 10 in the *C. savingyi* genes is the same length (67 bp) as exon 9 in human CYP3A4, and *C. intestinalis* exon 10 is just two codons longer. The *Ciona* gene exons 3, 4, 9, and 14 are only one or two codons different in length compared to counterpart codons in human CYP3A4.

This similarity in exon-intron structure is even more striking when compared to CYP6s and CYP9s, other invertebrate members of the CYP clan 3 (Supplementary material Table 4). The mosquito *Anopheles gambiae* possesses 29 CYP6 proteins and 8 CYP9s, and all contain 2 to 5 exons. The fruit fly *Drosophila melanogaster* contains 22 expressed CYP6 proteins (and three

CYP6 pseudogenes), and four expressed CYP9s (and one CYP9 pseudogene); all of these *Drosophila* CYPs contain two to five exons. In addition, the honeybee contains 22 CYP6 and 7 CYP9 genes, all of which contain 1-6 exons. (The gene structures are not known for *Homarus americanus* CYP45 (Genbank accession no. **AF065892**) and *Mercenaria mercenaria* CYP30 (Genbank accession no. **AF014795**), two other genes that cluster near the *Ciona* genes.) On the basis of known CYP gene exon-intron organization, the novel *Ciona* CYPs identified in this study are most closely related to vertebrate CYP3s rather than to insect CYP6s and CYP9s.

To better understand the possible functional evolution of these genes, we employed the program DIVERGE (Gu and Vander Velden, 2002), which measures the change in site-specific evolutionary rates by comparing these rates among subclades within a phylogenetic tree. We used DIVERGE to test the null hypothesis of no changes in site-specific evolutionary rates between the *Ciona* CYP subclades identified here and two vertebrate CYP3 subclades (mammal versus fish, Table 2). We found that the coefficient of evolutionary functional divergence (θ) is significantly different ($\theta = 0.33$, LRT = 14.44) between the *Ciona* CYPs and the mammalian CYP3As, as expected, but not significantly different between the *Ciona* CYPs and the fish CYP3s ($\theta = 0.13$, LRT = 0.68).

3.4. Gene expression patterns as derived from EST data

The four novel *C. intestinalis* CYP sequences were used to search the *C. intestinalis* EST/cDNA projects database (see Material and Method section). Table 3 shows the corresponding cDNA clusters and their DDBJ (Genbank/EMBL) accession number, the EST IDs, and the number of counts reflecting expression in different developmental stages or tissues. CI43 was only expressed in mature adult animals, whereas the other genes were expressed during different life stages (eggs, cleaving embryos, embryo mix, and adults). In addition, CI43 was only expressed in blood cells and CI19 only in the neural complex, whereas the other two genes were expressed in the neural complex, blood cells, heart, gonad, and digestive gland. From these EST counts it appears that CI85 is the most widely expressed of these four CYPs.

4. Discussion

To elucidate the possible origin of the gene lineage leading to the CYP3s in vertebrates, we used a profile-based HMM search for CYP3-related sequences in the predicted protein set of the tunicate *C. intestinalis*. This led to the identification of four CYP3A-like genes in *C. intestinalis*, three multi-exon genes and one single exon gene. Subsequently, homologs of these four genes were identified in *C. savignyi*. These results establish the occurrence of CYP clan-3 genes in the deuterostome line prior to vertebrates; the characteristics of these genes strongly suggest that they represent a tunicate CYP3 line. The multiplicity of these genes also indicates that diverse functions may be carried out by CYP3 genes in early-diverging chordates, as appears to be the case in vertebrates.

When a novel CYP3 is discovered, it is usually compared to human CYP3A4, the best known member of the CYP3A subfamily (Williams et al., 2004). A guiding principle for nomenclature has been a requirement of a shared amino acid identity of greater than 40% to be classified in the same family (Nelson et al., 1996). However, several issues arise with respect to this CYP nomenclature because of the evolutionary distance of *Ciona* from the other species in which CYP3s have been described. Although we used a targeted CYP3A HMM profile to search for novel CYP3A-related genes in the *Ciona* genome, the newly identified CYPs might not be considered CYP3s based on the criteria of sequence identity. However, Bayesian phylogenetic analyses show that the *Ciona* CYP genes do indeed group together with CYP3 sequences, and structural similarities lead us to suggest that these may be considered tunicate CYP3 genes.

The three multi-exon genes occur in tandem in the *C. intestinalis* genome. There is precedent for recombination between tandemly duplicated CYPs (Sinnott et al., 1990; Pascoe et al., 1992; Goldstone and Stegeman, in press), suggesting the possibility of gene conversion here. We observed no convincing evidence for gene recombination between *C. savignyi* genes, and only limited evidence for recombination between *C. intestinalis* genes, indicating that the phylogenetic relationships established by the Bayesian analysis are correct. However, given the overall high degree of conservation between the *C. intestinalis* genes it is difficult to distinguish between functional conservation and gene conversion.

Recent phylogenetic studies on the CYP3A subfamily have indicated that CYP3A diversity is the product of recent gene duplication events, leading us to suggest that the ancestral mammalian genome contained a single CYP3A gene (McArthur et al., 2003). However, while the CYP3A subfamily has a clear root between the fish subclade and the mammalian/diapsid subclade in that study, it was not possible to determine an appropriate phylogenetic root for the CYP3 family as a whole, indicating that CYP3 origins may be obscured by incomplete taxonomic sampling (A. McArthur, Marine Biological Laboratory, Woods Hole, MA, USA, personal communication). While our findings do not necessarily provide that root, they do indicate a common evolutionary history for these tunicate sequences and the vertebrate CYP3s. With the availability of a genome sequence for the sea urchin *Strongylocentrotus purpuratus*, it will be interesting to look for similar genes in echinoderms, which phylogenetically are deuterostomes diverging even earlier than the tunicates. Based on the current understanding, we would expect CYP3-like genes in *S. purpuratus*.

It has been suggested that all CYP3 genes arose from a common ancestral gene lineage that diverged between 800 and 1100 mya, before the deuterostome-protostome split (Nelson, 1998; Williams et al., 2004). The very strong support (Bayes Factor = 15.6) for the positioning of the *Mercenaria* CYP30 with the tunicate and vertebrate CYP3 genes is consistent with this hypothesis. The relative positioning of the ecdysozoans (e.g., insects) and lophotrochozoans (e.g., bivalves) within metazoan phylogeny remains the subject of active debate, both based on molecular and morphological approaches (Glenner et al., 2004). Nevertheless, it is possible that a CYP3 lineage in a bilaterian ancestor diverged into CYP6s and CYP9s in the ecdysozoa, CYP30 in the lophotrochozoa, and the CYP3s in the deuterostomata. The reason for the positioning of the CYP5 genes basal to the CYP3s remains unclear, as it appears that the CYP5s may have evolved from an ancestral CYP3, but the somewhat divergent functionality of CYP5 genes (thromboxane synthase) may only be present in vertebrates.

In this present study, the use of higher order "clans" as proposed by Nelson (Nelson, 1998) helps to describe relationships between more distantly related CYPs, occupying distinct gene families. The CYP clan 3 contains the CYP3s, CYP5s, CYP6s CYP9s, as well as the *Ciona* genes and selected others. However, based on the comparison of gene exon-intron organization shown in Figure 4, the *Ciona* CYPs appear to be more closely related to the CYP3s than to other CYP families in clan 3. Thus, as indicated above, we suggest that these early diverging representatives of the gene line leading to the vertebrate CYP3s are tunicate CYP3s, very similar to the vertebrate CYP3s. Unfortunately, these relationships cannot be addressed adequately within the current nomenclature. There is need for a nomenclature that, going forward, employs principles (Thornton and DeSalle, 2000) that accommodate evolutionary distances greater than just among vertebrates, so as not to obscure distant orthologous relationships.

In addition to the multi-exon genes, two of the *Ciona* CYPs identified in this study, CI19 and CSV A, contained no introns. Other single exon CYP genes have been predicted in *C. intestinalis* (JGI accession numbers are ci0100152549 and ci0100154611, two members of the CYP 2 clan), and at least one of these has a single exon ortholog in *C. savignyi*, i.e., ci0100154611. To the best of our knowledge, the occurrence and proportion of single exon CYPs within presently sequenced genomes has not been studied. However, a number of eukaryotic genomes contain a substantial number of single exon genes (an estimated 12% of genes in the human genome are single exon genes) and their proportion seems to be related to gene count and gene density (Sakharkar et al., 2004). The biological role of these single exon genes in genomes of eukaryotes is not completely understood, although their mechanism of origin is most likely re-insertion (Brosius, 1999). Re-insertion is caused by reverse transcription of mRNA and insertion of the processed gene into the genome producing an intronless duplicate. As orthologous single-exon genes were identified in both species in this study, we conclude that re-insertion would have occurred before speciation between *C. savignyi* and *C. intestinalis*. Many re-inserted genes are silent or devoid of meaningful open reading frames (Brosius, 1999). However, the *C. intestinalis* single exon CYP is expressed. Since the activity of an inserted retrogene depends on the presence or activation of an upstream promoter, and the distance of this promoter to the newly inserted retrogene could be large, future studies are necessary to determine if the expression of these single exon CYPs in *Ciona* differs from the multi-exon CYP genes.

Our identification of CYP3-like proteins in *Ciona* should also stimulate ongoing research on pin-pointing candidate regulatory motifs (and their putative binding factors) in the promoter region and their importance for CYP3 regulation. It will be interesting to determine whether a pregnane x receptor homolog occurs and is involved in regulation of these CYP3-like genes in urochordates. Because it is assumed that sequence conservation implies functional importance, comparative genomics could further help to identify regulatory motifs in CYP genes. It is important to note that intergenic distances were small in the CYP cluster found on chromosome 11, and as such CI50, CI85, and CI43 might be co-expressed. The existence of prokaryote-like operons within eukaryotic genomes has been demonstrated in the appendicularian *Oikopleura dioica*, where processing of polycistronic transcripts has been demonstrated (Ganot et al., 2004). However, current expression data suggest differential regulation of the novel *Ciona* genes.

As part of our study, we searched the *C. intestinalis* cDNA/EST resource (Satou et al., 2002; 2003) for clues on the expression profile of the *Ciona* CYPs. As shown in Table 3, differences were observed in the expression of the four novel CYP genes in *C. intestinalis*. While the results indicate tissue selectivity in expression of the various genes, the biological significance of these expression differences is not known. Some vertebrate CYP3As also are widely expressed, for example in liver, brain, kidney, testis, gonad, stomach, and heart of rainbow trout (Lee et al., 1998). The *Ciona* genes also are expressed at different stages of development indicating a potential stage-specific functional role. Previous studies have demonstrated the developmental expression of multiple CYP3As (Stevens et al., 2003; Aleksa et al., 2005). Thus, CYP3A4 and CYP3A7 are expressed specifically in adult and fetal human livers, respectively (Komori et al., 1990), and distinct CYP3As are differentially expressed in embryonic and adult fish (Kullman and Hinton, 2001).

In summary, four novel CYP clan-3 genes were identified and are expressed in *C. intestinalis,* and homologues of these were identified in the *C. savignyi* genome. These include orthologous single exon genes identified in both *C. intestinalis* and *C. savignyi*, likely caused by a retro-transposition event prior to speciation within the genus *Ciona*. Our study provides convincing evidence that these CYPs identified in *Ciona* are early diverging representatives of the line leading to the vertebrate CYP3 genes. Vertebrate CYP3s as well as the arthropod CYP6s and CYP9s are all involved in steroid hormone metabolism and/or the metabolism of various xenobiotics (Snyder, 1998 and references therein). The functional characterization of these new *Ciona* CYPs should indicate whether tunicate CYP3-like enzymes play similar roles in early chordate biochemistry and physiology, and could identify evolutionarily conserved and thus possible original (endogenous) functions.

Acknowledgements

These studies were supported in part by NIH grant 2-P42-ES07381 to J.J. Stegeman. Tim Verslycke was supported by the Postdoctoral Scholar Program at the Woods Hole Oceanographic Institution, with funding from the Ocean Life Institute and by a Fellowship of the Belgian American Educational Foundation. Jared Goldstone was supported by a Ruth Kirschstein National Research Service Award (NIH 5F32ES 012794). Computational support was provided by the WM Keck Ecological and Evolutionary Genetics Facility at the Marine Biological Laboratory. We thank David Nelson for helpful discussion of CYP classification in general, and of the CYP clan concept in relation to the placement of the *Ciona* genes, Ann Tarrant and Bruce Woodin for technical advice, and Andrew McArthur for provision of his earlier CYP3 alignment.

References

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

Aleksa, K., Matsell, D., Krausz, K., Gelboin, H., Ito, S., Koren, G., 2005. Cytochrome P450 3A and 2B6 in the developing kidney: implications for ifosfamide nephrotoxicity. Pediatr. Nephrol. 20, 872-885.

Bresolin, T., Rebelo, M.D., Bainy, A.C.D., 2005. Expression of PXR, CYP3A and MDR1 genes in liver of zebrafish. Comp. Biochem. Physiol. C 140, 403-407

Brosius, J., 1999. Many G-protein-coupled receptors are encoded by retrogenes. Trends Genet. 8, 304-305.

Brown D.J., Clark G.C., Van Beneden R.J., 1998. A new cytochrome P450 (CYP30) family identified in the clam, *Mercenaria mercenaria*. Comp. Biochem. Physiol. C 121, 351-360.

Celander, M., Buhler, D.R., Forlin, L., Goksoyr, A., Miranda, C.L., Woodin, B.R., Stegeman, J.J., 1996. Immunochemical relationships of cytochrome P4503A-like proteins in teleost fish. Fish Physiol. Biochem. 15, 323-332.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., Harafuji, N., Hastings, K.E.M., Ho, I., Hotta, K., Huang, W., Kawashima, T., Lemaire, P., Martinez, D., Meinertzhagen, I.A., Necula, S., Nonaka, M., Putnam, N., Rash, S., Saiga, H., Satake, M., Terry, A., Yamada, L., Wang, H.G., Awazu, S., Azumi, K., Boore, J., Branno, M., Chin-bow, S., DeSantis, R., Doyle, S., Francino, P., Keys, D.N., Haga, S., Hayashi, H., Hino, K., Imai, K.S., Inaba, K., Kano, S., Kobayashi, K., Kobayashi, M., Lee, B.I., Makabe, K.W., Manohar, C., Matassi, G., Medina, M., Mochizuki, Y., Mount, S., Morishita, T., Miura, S., Nakayama, A., Nishizaka, S., Nomoto, H., Ohta, F., Oishi, K., Rigoutsos, I., Sano, M., Sasaki, A., Sasakura, Y., Shoguchi, E., Shin-I, T., Spagnuolo, A., Stainier, D., Suzuki, M.M., Tassy, O., Takatori, N., Tokuoka, M., Yagi, K., Yoshizaki, F., Wada, S., Zhang, C., Hyatt, P.D., Larimer, F., Detter, C., Doggett, N., Glavina, T., Hawkins, T., Richardson, P., Lucas, S., Kohara, Y., Levine, M., Satoh, N., Rokhsar, D.S., 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. Science 298, 2157-2167.

Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H., 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965-968.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755-763.

Finta, C., Zaphiropoulos, P.G., 2000. The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons. Gene 260, 12-23.

Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D., Thompson, E.M., 2004. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. Mol. Cell. Biol. 24, 7795-7805.

Glenner, H., Hansen, A.J., Sørensen M.V., Ronquist, F., Huelsenbeck, J.P., Willerslev, E., 2004. Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach. Curr. Biol. 14, 1644-1649.

Goldstone, H.M.H., Stegeman, J.J., in press. Evidence of a single gene duplication and gene conversion in tetrapod CYP1As. J. Mol. Evol.

Gonzalez, F.J., 1990. Molecular genetics of the P-450 superfamily. Pharm. Ther. 45, 1-38.

Graham, A., 2004. Evolution and development: rise of the little squirts. Curr. Biol. 14, R956-R958.

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

Guengerich, F.P., 1991. Reactions and significance of cytochrome P-450 enzymes. J. Biol.

Chem. 266,10019-10022.

Guengerich, F.P., 1999. Cytochrome P-450 3A4: Regulation and role in drug metabolism. Annu. Rev. Pharmacol. Toxicol. 39, 1-17.

Guindon S, Gascuel O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52(5):696-704.

Hashimoto, H., Toide, K., Kitamura, R., Fujita, M., Tagawa, S., Itoh, S., Kamataki, T., 1993. Gene structure of CYP3A4, an adult-specific form of cytochrome P-450 in human livers, and its transcriptional control. Eur. J. Biochem. 218, 585-595.

Hegelund, T., Celander, M.C., 2003. Hepatic versus extrahepatic expression of CYP3A30 and CYP3A56 in adult killifish (*Fundulus heteroclitus*). Aquat. Toxicol. 64, 277-291.

Huelsenbeck, J.P., Imennov, N.S., 2002. Geographic origin of human mitochondrial DNA: accommodating phylogenetic uncertainty and model comparison. Syst Biol. 51, 155-65.

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. J. Am. Stat. Assoc. 90, 773-795.

Komori, M., Nishio, K., Kitada, M., Shiramatsu, K., Muroya, K., Soma, M., Nagashima, K., Kamataki, T., 1990. Fetus-specific expression of a form of cytochrome P-450 in human livers. Biochemistry 29, 4430-4433.

Kullman, S.W., Hinton, D.E., 2001. Identification, characterization, and ontogeny of a second cytochrome P450 3A gene from the fresh water teleost medaka (*Oryzias latipes*). Mol. Reprod. Dev. 58, 149-158.

Lee, S.J., Wang-Buhler, J.L., Cok, I., Yu, T.S., Yang, Y.H., Miranda, C.L., Lech, J., Buhler, D.R., 1998. Cloning, sequencing, and tissue expression of CYP3A27, a new member of the CYP3A subfamily from embryonic and adult rainbow trout livers. Arch. Biochem. Biophys. 360, 53-61.

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

Maurel, P., 1996. The CYP3 Family. In: Ioannides, C. (Ed.), Cytochromes P450 metabolic and toxicological aspects. CRC Press, Boca Raton, FL, pp. 241-270.

McArthur, A.G., Hegelund, T., Cox, R.L., Stegeman, J.J., Liljenberg, M., Olsson, U., Sundberg, P., Celander, M.C., 2003. Phylogenetic analysis of the cytochrome P450 3 (CYP3) gene family. J. Mol. Evol. 57, 1-12.

Nebert, D.W., Gonzalez, F.J., 1987. P450 genes: Structure, evolution, and regulation. Annu. Rev. Biochem. 56, 945-993.

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

Nelson, D.R., 1998. Metazoan cytochrome P450 evolution. Comp Biochem Physiol C 121, 15-22.

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

Ranson, H., Nikou, D., Hutchinson, M., Wang, X., Roth, C.W., Hemingway, J., Collins, F.H., 2002. Molecular analysis of multiple cytochromes P450 genes from the malaria vector, *Anopheles gambiae*. Insect Mol. Biol. 11, 409-418.

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

Sakharkar, M.K., Chow, V.T.K., Chaturvedi, I., Mathura, V.S., Shapshak, P., Kangueane, P., 2004. A report on single exon genes (SEG) in eukaryotes. Frontiers in Bioscience 9, 3262-3267.

Satou, Y., Takatori, N., Fujiwara, S., Nishikata, T., Saiga, H., Kusakabe, T., Shin-I, T., Kohara, Y., Satoh, N., 2002. *Ciona intestinalis* cDNA projects: expressed sequence tag analyses and gene expression profiles during embryogenesis. Gene 287, 83-96.

Satou, Y., Kawashima, T., Kohara, Y., Satoh, N., 2003. Large scale EST analyses in *Ciona intestinalis* - Its application as Northern blot analyses. Dev. Genes Evol. 213, 314-318.

Schaeffer, B., 1987. Deuterstome monophyly and phylogeny. Evol. Biol. 21, 179-235.

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

Snyder M.J., 1998. Identification of a new cytochrome P450 family, CYP45, from the lobster, *Homarus americanus*, amd expression following hormone and xenobiotic exposures. Arch. Biochem. Biophys. 2, 271-276.

Solovyev, V.V., Salamov, A.A., 1999. INFOGENE: A database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. Nucleic Acids Res. 27, 248-250.

Stegeman, J.J., Livingstone, D.R., 1998. Forms and function of cytochrome P450. Comp. Biochem. Physiol. C 121,1-3.

Stevens, J.C., Hines, R.N., Gu, C.G., Koukouritaki, S.B., Manro, J.R., Tandler, P.J., Zaya, M.J., 2003. Developmental expression of the major human hepatic CYP3A enzymes. J. Pharmacol. Exp. Therapeut. 307, 573-582.

Suchard, M.A., Weiss, R.E., Sinsheimer, J.S., 2005. Models for estimating bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61, 665-73.

Swofford, D.L., 2003. PAUP*: Phylogenetic analysis using parsimony (* and other methods) v4b10. Sinauer, Sunderland, Massachusetts.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX-Windows interface: Flexible strategies for multiple sequence allignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876-4882.

Thornton, J.W., DeSalle, R., 2000. Gene family evolution and homology: Genomics meets phylogenetics. Annu. Rev. Genomics Hum. Genet. 1, 41-73.

Thummel, K.E., Wilkinson, G.R., 1998. In vitro and in vivo drug interactions involving human CYP3A. Ann. Rev. Pharmacol. Toxicol. 38, 389-430.

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

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, 691–699.

Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. http://ceb.csit.fsu.edu/awty.

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

Zeng, L.Y., Swalla, B.J., 2005. Molecular phylogeny of the protochordates: chordate evolution. Can. J. Zool. 83, 24-33.